UNIVERSITEIT GENT

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Essays on Data Augmentation: the Value of Additional Information

Philippe Baecke

2012

Advisor:

Prof. Dr. Dirk Van den Poel

Dissertation submitted to the Faculty of Economics and Business Administration, Ghent University, in fulfillment of the requirements for the degree of Doctor in Applied Economics

PhD Series - Ghent University, September 2012
Faculty of Economics and Business Administration
http://www.FEB.UGent.be

DOCTORAL COMMITTEE

Prof. dr. Marc De Clercq
(Ghent University, Dean)

Prof. dr. Patrick Van Kenhove
(Ghent University, Academic Secretary)

Prof. dr. Dirk Van den Poel
(Ghent University, Advisor)

Prof. dr. Anita Prinzie
(Ghent University)

Prof. dr. Bart Baesens
(University of Leuven)

Prof. dr. Martin Natter
(University of Frankfurt)

TABLE OF CONTENTS

ACKNOWLEDGEMENTS (DANKWOORD)

In tegenstelling tot wat mijn collega Dries schrijft in zijn doctoraat, wil ik dit dankwoord beginnen met de woorden: “Een doctoraat schrijf je niet alleen”. Ondanks het feit dat het uitvoeren van academisch onderzoek vaak een individualistische bezigheid is, kan de steun van een aantal personen cruciaal zijn in het al dan niet slagen van zulk project. Daarom wil ik hier een aantal mensen bedanken.

Om te beginnen wil ik graag de leden van mijn leescommissie bedanken. Prof. dr. Anita Prinzie, prof. dr. Bart Baesens, prof. dr. Martin Natter. Bedankt voor de tijd die jullie hebben geïnvesteerd in het kritisch nalezen van mijn doctoraat. Jullie opmerkingen hebben het niveau van dit doctoraat zeker opgetild en bovendien hebben jullie mij geïnspireerd met nieuwe ideeën voor toekomstig onderzoek.

Verder had ik graag de decaan van deze faculteit, prof. dr. Marc DeClerq, bedankt. Ik heb genoegen gehad om 10 jaar op deze faculteit aanwezig te mogen zijn. 5 jaar als student en nog eens 5 jaar als doctoraatstudent. In beide hoedanigheden heb ik deze omgeving als heel leerrijk, maar ook als heel aangenaam ervaren.

Iemand die ik zeker ook wil vermelden is prof dr. Patrick van Kenhove (academisch secretaris). Op onze vakgroep heerst een heel sociale werksfeer en daar heb jij, als vakgroepvoorzitter, ongetwijfeld ook een invloed op. Jij hebt me de passie voor marketing overgebracht. Het is bewonderenswaardig met hoeveel overgave je les kan geven over marktonderzoek. Ik kan me nog goed herinneren dat ik als student van jou de opdracht kreeg om AC Nielsen cijfers te analyseren over spaghetti sauzen. Dat was het moment waar ik dacht dat dit misschien wel iets voor mij zou zijn, waardoor ik uiteindelijk heb beslist om de Master na Master in Marketing Analysis te volgen.

Dit brengt me tot de persoon die dit doctoraat heeft mogelijk gemaakt voor mij: mijn promotor, prof. dr. Dirk Van den Poel. In de Master na Master in Marketing Analysis heb je me ervan overtuigd dat ik ook in Marketing mijn analytische capaciteiten ten volle kan benutten. Ik ben je heel erg dankbaar voor het de kans die je me gegeven hebt om na deze master te doctoreren.

Tijdens mijn onderzoek heb je mij steeds voldoende vrijheid gegeven, waardoor ik mezelf ten volle heb kunnen ontplooien en me kon focussen op de dingen die me echt interesseren. Echter soms kon ik wel eens vloeken wanneer je het bureau binnen komt en iets vraagt in de aard van: “Philippe, zou je die paper tegen volgende week kunnen afwerken”, terwijl ik op dat moment nog maar aan de introductie zat. Maar dit gaf me de nodige boost die mijn productiviteit de hoogte in jaagde. Ondanks het feit dat onze bureaus nu op een iets verder van elkaar liggen, hoop ik dat we deze samenwerking nog lang mogen verder zetten. Verder wil ik je bedanken voor de plezante momenten op de vele conferenties de we samen hebben gevolgd, op soms wel exotische locaties zoals Las Vegas, Vancouver en Miami. Daarnaast wil ik je ook bedanken voor de vlotte samenwerking in de verschillende bedrijfsprojecten.

Een doctoraat kost namelijk ook geld, daarom wil ik hier graag ook de bedrijven vermelden die mijn doctoraat mede gefinancierd hebben. Ten eerste, WDM Belgium, en meer specifiek Alain Pletinckx, die hier jammer genoeg niet meer kan zijn. Deze persoon heeft me namelijk het belang laten inzien van praktisch toepasbaar academisch onderzoek. Ten tweede, Belgian Icecream Group, ook hen wil ik bedanken voor het vertrouwen in onze analytische modellen, die op dat moment vrij nieuw waren voor hen.

En nu zijn we aangekomen bij de collega’s. Zoals ik al zei is het werken aan een doctoraat een vrij individualistische bezigheid. Toch heb ik er bewust voor gekozen om niet echt veel van thuis uit te werken. Dit is voornamelijk omdat we binnen de vakgroep marketing een groep met heel toffe collega’s hebben. Ik was misschien niet de persoon die overal telkens de deur plat liep, daar hadden we dries namelijk voor, maar ik heb toch genoten van de social fridays, de plezante babbels, de steeds goed georganiseerde vakgroepactiviteiten, de facebook anecdotes, en de feestjes met de collega’s. Aangezien de Vlerick nu niet echt ver is, hoop ik jullie toch nog steeds regelmatig eens terug te zien.

Het is onmogelijk om iedereen te bedanken, maar toch had ik hier graag een paar collega’s of ex-collega’s een speciale vermelding gegeven. Ik had graag begonnen met “la mama” van de vakgroep: Karin. Het is ongelofelijk hoe je altijd iedereen met elke administratieve aangelegenheid kunt helpen. Zonder jou had ik nu waarschijnlijk nog steeds geen zaal om mijn doctoraatsverdediging te houden.

Ook Dauwe had ik graag bedankt. Niet alleen omdat hij dit vroeg, maar ook omwille van de vlotte samenwerking in het ijsboerke project. Ik hoop dat we hier nog een mooie paper kunnen uithalen. Daarnaast wil ik je ook bedanken voor de pure entertainment die je in onze bureau binnen bracht. Het was telkens heel amusant om het doctoraatswerk af te wissel met een verhaaltje uit "het leven van Dauwe".

Coussy, jouw wil ik ook bedanken. Niet alleen voor de samenwerking in het vak marketing information systems, maar je hebt me ook geleerd iets kritischer te kijken naar zowel de academische wereld als de bedrijfswereld. Voorst heb ik veel bewondering hoe je er in slaagt om zo snel, zoveel academische output te creëren: een echte paper machine.

Bocky, ik heb het genoegen gehad om drie jaar samen met jou op de bureau te zitten. We hebben altijd heel goed kunnen opschieten. Ik wil je bedanken voor de waardevolle feedback op mijn onderzoek, de toffe babbels, de plezante tijden op conferentie en de vele vendridi sports. Ik ben er van overtuigd dat we dit nog wel vaak zullen verder zetten. Aangezien we binnenkort niet ver meer van elkaar wonen zou ik het misschien wel eens durven wagen om samen met jou nog eens op mijn fiets te kruipen, als die nog niet volledig verroest is tenminste.

Nu zijn we aan de huidige bureaugenoten aangekomen. William, bedankt om me onder te dompelen in de wondere wereld van social media en smatphones. In beschouw mezelf eerder als een follower in die zaken, maar ik moet toegeven dat zo een HTC wel nuttiger is dan ik oorspronkelijk dacht. Ik ben er ook van overtuigd dat je facebook initiatieven een ware meerwaarde geweest zijn voor de vakgroep.

Michi, de Lucky Luke van de foto's. Jij trekt sneller en meer foto's dan al de rest van de collega's samen. Dit maakte het heel handig om tijdens een periode van afwezigheid toch de gebeurtenissen op de vakgroep te kunnen opvolgen. Verder wil ik je bedanken voor de dagelijkse portie vitamientjes die me net die extra 5% percent productiviteit gaven om mijn doctoraat tot een goed eind te brengen. Ik kijk nu al uit naar jouw verdediging, zodat ik eindelijk weet wat je met die facebook data van plan bent.

Tot slot, Benny, jij hebt me zien komen op de vakgroep en zal me nu ook zien vertrekken. 5 jaar hebben we dezelfde bureau gedeeld. We hebben veel geanimeerde discussies gehad. Soms over onderzoek of politiek, maar minstens evenveel over andere zaken. Deze laatste vaak geïnitieerd met de woorden “Stel dat ...”. Aangezien ik al vaak eens lach met onze West-Vlaamse medemens zou je het misschien niet onmiddellijk denken, maar ik heb veel respect voor jou. Als onderzoeker vind ik het heel knap hoe je jezelf zo hebt kunnen specialiseren in een moeilijk domein: Bayesian statistics dan nog wel. Wanneer ik weer eens 3 bladzijden vol met handgeschreven formules zie liggen op je bureau zeg ik al snel dat je weer statisch aan het *********** bent, maar dit is eerder uit bewondering hoe sterk je jezelf hierin hebt ontwikkeld. Ook als persoon vind ik het buitengewoon hoe sociaal jij wel bent. Dit heeft als voordeel dat ik niet eens van mijn bureau dien te komen en toch op de hoogte ben van de laatste roddels binnen de vakgroep. Ondanks het feit dat we niet dezelfde type van personen zijn, kunnen we het toch heel goed met elkaar vinden. Ik hoop dan ook dat we in de toekomst nog regelmatig eens zullen lunchen, een pint gaan drinken of zelfs een stapke zetten.

Naast de collega’s had ik graag nog de vrienden buiten de universiteit bedankt. Dit zijn dan voornamelijk de voetbal vrienden: Kenny Cola, The Wall Jensen, Tomaba La Bomba, Nicopolonaise, Tho International, Captain Freddy, Mr 50% Bramme, Razorblade Nick, Brutus en de gebroeders Gallez. Jullie zorgen steeds voor de nodige afleiding en ontspanning die ik nodig heb. Vele van jullie ken ik al van toen we nog als klein mannen aan het shotten waren achter de sporthal of op KVV. Ondanks het feit dat we voetballend niet veel progressie meer maken, denk ik dat het qua ploegsfeer niet beter kan. MVC is dan ook meer dan voetballen alleen.

Tot slot wil ik mijn broers en mijn ouders bedanken. Xavier, hopelijk sta jij binnen een aantal jaar hier ook met een doctoraat in de geschiedenis. Christophe en Laurent, jullie zijn steeds de eerste die ik lastig val wanneer ik iemand nodig heb: de reparatie van een mijn auto, de inrichting van mijn huis of het vullen van een container vol met steen. Heel erg bedankt dat jullie er telkens voor mij staan. Last, but not least, Moemoe, je zal waarschijnlijk heel fier zijn op wat ik hier doe. Misschien zelfs nog meer dan ikzelf. “We zijn wij maar gewone mensen” zeg je dan altijd. Bij deze wil ik je toch nog eens een speciale vermelding geven. Vooral voor de manier waarop je ons opgevoed hebt. We hebben steeds de vrijheid gekregen om te doen en laten wat we willen, maar toch heb je ons altijd gesteund in wat we deden. Ook wanneer ik een huis kocht dat helemaal vervallen was, ben jij de persoon die me steeds komt helpen. Bedankt voor alles!

Philippe Baecke, 21 september 2012

DUTCH SUMMARY

(NEDERLANDSTALIGE SAMENVATTING)

De laatste twee decennia is er binnen marketing een opvallende evolutie waar te nemen van traditionele massa marketing naar klantenrelatie beheer (*Customer Relationship Management* of *CRM)*. Hierbij worden intelligente informatiesystemen gebruikt om de communicatie tussen bedrijf en klant te personaliseren met het oog om een lange termijn relatie met deze klanten op te bouwen. In het algemeen kan men deze analytische CRM modellen op twee manieren optimaliseren. Enerzijds kan men focussen op het verbeteren de statistische modellen die gebruikt worden om klanten op een individuele manier te behandelen. Anderzijds kan men proberen om de kwaliteit van de datasets waarop deze modellen zijn gebaseerd te verbeteren. Deze doctoraatsverhandeling concentreert zich voornamelijk op dit laatste aspect. In vier verschillende studies word besproken hoe alternatieve vormen van data op een creatieve manier bestaande datasets kunnen verrijken. Dit gebeurt met de bedoeling om de voorspellende prestaties van CRM modellen die hierop gebaseerd zijn te verbeteren.

De eerste studie richt zich op het opnemen van gesurveilleerde data in CRM modellen. Traditioneel worden CRM modellen toegepast op transactionele data verkregen van een grote hoeveelheid klanten. Echter, naast koopgedrag speelt ook de attitude van een persoon ten opzichte van een bepaald product een belangrijke rol om de bepalen hoe waardevol een specifieke klant is. Echter, deze data kan slechts voor een beperkte groep van respondenten via vragenlijsten worden verzameld. In combinatie met een dataset van een externe dataverkoper, presenteert deze studie een methodologie, gebaseerd op *Random Forests data mining* technieken, om deze te extrapoleren naar alle respondenten van de externe dataverkoper. Op deze manier werden in 26 productcategorieën *spending pleasure* variabelen ontwikkeld waarvan hun meerwaarde bewezen werd in een toepassing waarbij potentiële nieuwe klanten werden geïdentificeerd voor een magazine uitgever.

De tweede studie merkt op dat het aankoopgedrag van een consument vaak wordt beïnvloed door situationele variabelen. Echter, hier wordt in de CRM literatuur weinig aandacht aan besteed.

Daarom onderzoekt deze studie in een *home vending* industrie de meerwaarde van drie types situationele variabelen, namelijk fysische omgevingsfactoren (het weer), tijdsperspectief (het moment van de dag) en sociale omgevingsfactoren (de verkoper aanwezig). Hieruit blijkt dat het opnemen van elk van deze situationele variabelen de voorspellingen met betrekking tot het aankoopgedrag van de consument verbeteren.

De derde studie is gefocust op de incorporatie van geografische informatie in traditionele CRM modellen. Deze traditionele CRM modellen veronderstellen namelijk vaak dat het aankoopgedrag van de consument volledig onafhankelijk gebeurt, terwijl dit in realiteit regelmatig wordt beïnvloed door derden. Als gevolg, het mee in rekening brengen van deze informatie kan helpen om potentiële klanten beter te identificeren. Deze studie toont aan dat het diepteniveau waarop geografisch gecorreleerd gedrag wordt gemeten kan een belangrijke invloed hebben op de prestatie van het model. Verder wijst deze studie op het feit dat geografische gecorreleerd gedrag verschillende oorzaken kan hebben, elk gemeten op een optimaal diepteniveau. Daarom is het nuttig om, indien er voldoende data beschikbaar is, meerdere diepteniveau gelijktijdig op te nemen.

De vierde studie spits zich ook toe op het belang van geografische informatie om beter potentiële klanten te identificeren. Echter, in deze studie werden twee statistische technieken vergeleken die in staat zijn om geografisch gecorreleerd gedrag op te nemen, namelijk een hiërarchische en een autoregressieve techniek. Bovendien werd het belang van geografische informatie onderzocht voor 25 verschillende merken en producten. De resultaten gaven weer dat wanneer een variabele wordt gebruikt die respondenten opsplitst in mutueel exclusieve groepen een hiërarchische techniek te verkiezen is. Voorst bleek ook dat het belang van geografische informatie in CRM modellen sterk toeneemt wanneer het gaat om publiek geconsumeerde duurzame goederen.

CHAPTER I:

GENERAL INTRODUCTION

CHAPTER I:
GENERAL INTRODUCTION

1. Introduction

Within the field of marketing there has been an important evolution in the way how companies try to target their customers. Whereas mass marketing has been an established way of communication for several decades, marketers gradually tried to segment customers with similar characteristics into specific groups. By this, companies are more able to adopt their marketing actions to the needs and demands of specific groups instead of approaching all customers in the same way. For example, in the 1960s, marketers realized that ZIP codes could be useful to segment customers geographically. This is because surrounding people tend to influence each other's behavior and attitudes, but also because people with similar characteristics and interests have the tendency to cluster together (Baier, 1967).

Since the 1980s, the increase in computational power has brought a revolution in the tools available for customer targeting. This resulted in a new fast-growing domain within marketing, called database marketing, in which individual customer characteristics are recorded in electronic databases (Petrison, Blattberg, & Wang, 1993). This should help companies to learn more about their customers and be able to target them more effectively. The technological evolution made it even possible to evolve towards more individualized marketing including personalized communications, products and services that meet each customer's needs in order to build a long-term relationship. This should eventually translate into substantial profits. During the 1990s, in an increasingly competitive environment, customer relationship management (CRM) has emerged as an important discipline within marketing. This discipline points to a new dimension for competition, namely the relationship dimension. In a saturated market with a lot of competitors that are doing the same thing, a competitive product or a reasonable price will not give a company a long-term advantage because competitors quickly react with new products or prices. However, using CRM strategies, companies try to influence how customers feel about a company, which is a much more sustainable advantage. In fact this kind of relationship marketing dates back to the earliest days of direct marketing. Smaller companies with a limited number of customers were able to address their customers individually, but when they began growing, it became too laborious and too costly to communicate with consumers on a one-on-one basis (Petrison et al., 1993). However, the exponential increase of computational power, the drop in data warehousing costs and the rise of the internet have resurfaced the idea of personalized communication, but now on a larger scale.

2. Customer Relationship Management

In general, CRM can be subdivided into operational and analytical CRM (Teo, Devadoss, & Pan, 2006). While operational CRM is focused on the automation of business processes (e.g. sales force automation and call centers), this dissertation is situated in the field of analytical customer relationship management (aCRM). Analytical CRM describes the component of CRM that analyzes data collected about customers to create a deeper understanding of its customers' behaviors. Eventually, this should result in better marketing decision. To this end, data mining techniques are frequently implemented to transform a large amount of unstructured data into valuable knowledge that can be used to support and forecast the effect of several marketing strategies (Ngai, Xiu, & Chau, 2009).

Analytical CRM can support marketing decision makers to effectively allocate resources across the different stages of the customer lifecycle, namely customer acquisition, customer development and customer retention (Kamakura et al., 2005).

Firstly, customer acquisition involves identifying and attracting these prospects that are most likely to become a customer or most profitable to the company. Such customer identification strategies can be executed using customer segmentation techniques or target customer analyses. In customer segmentation, customers are divided into smaller groups of customers with similar characteristics. This dissertation is more focused on target customer analyses in which data mining techniques are used to uncover the effect of customers' underlying characteristics on their purchasing behavior and preferences. This makes it possible to rank potential customers based on their expected value for the company. Next, only these prospects can be targeted for which this expected value is high enough (e.g. Buckinx, Moons, Van den Poel, & Wets, 2004).

Secondly, customer development strategies aim to increase customer transaction intensity, customer transaction value and customer profitability. Common activities within customer development are customer lifetime value analysis, market basket analysis and up/cross-selling. Customer lifetime value analyses are applied to estimate the value of a customer based on the expected future cash-flow this customer will generate (e.g. Baesens et al., 2004). Also market basket analyses can improve customer profitability by revealing regularities in the purchasing behavior of the customer (e.g. Chen, Tang, Shen, & Hu, 2005). These analyses can provide insights about which products customers buy, when they are purchased and in which sequence they are bought. This information can then be used to improve decisions about sales promotions, store designs, loyalty programs, but also to support cross-selling and up-selling strategies. Up-selling is focused on extending the demand of customers within the same product category, while cross-selling is involved with selling additional products or services to existing customers in order to extend their product portfolio (e.g. Prinzie & Van den Poel, 2007).

Thirdly, within CRM, most research has been devoted to customer retention. Customer retention strategies are focused on improving customer satisfaction and prolonging the relationship with the company. In this context, predictive techniques are often used to identify those customers who have a high probability to churn (Van den Poel & Larivière, 2004). This should help marketing decision makers to set up effective marketing action in order to increase the loyalty of these customers. Many studies have proven the importance of such loyalty programs. Especially since a strong link can be identified between customer retention and profitability. Since customer profitability increases over time, even a small improvement in customer retention can have a great impact on the firm’s total profitability (Gupta, Lehmann, & Stuart, 2012; Reichheld & Sasser, 1990). Closely related with customer retention is customer reactivation which tries to reactivate “sleeping” or reacquire lapsed customers (Thomas, Blattberg, & Fox, 2004).

3. Data augmentation in Customer Relationship Management

Plenty of researchers and business practitioners try to improve such aCRM models. In general, this can be done in two ways. Firstly, by improving the data mining techniques used to translate data into useful information for marketing actions, and secondly, by improving the data on which these data mining techniques are based.

Most of the academic research in the field of aCRM has focused on the development and comparison of new statistical or machine learning techniques. Among these techniques, a division can be made between unsupervised learning techniques and supervised learning techniques. Unsupervised learning techniques are frequently used to segment the market into different clusters that are internally homogeneous and mutually heterogeneous (Hung & Tsai, 2008). However, this dissertation is more situated in the field of supervised learning that tries to predict future customer behavior. As a result of the large amount of research in this field, database marketing techniques have evolved from traditional RFM models (based on recency, frequency and monetary value of customer purchases) over statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression (Bult & Wansbeek, 1995; Mccarty & Hastak, 2007) towards more advance machine learning techniques, such as support vector machines and neural networks (Shin & Cho, 2006; Zahavi & Levin, 1997) . Recently, more and more ensemble models are introduced, which combines multiple models to obtain better predictive performance (De Bock & Van den Poel, 2011).

Although this dissertation will use some advanced algorithms, such as random forests, or statistical techniques that are less commonly used for predictive purposes, such as generalized linear mixed models and autologistic regression models, the main focus of this dissertation will be on data augmentation. Besides the data mining technique used, also the quality of the database can have an important influence on the predictive performance of a CRM model. The database of a company, on which all data mining techniques are based on, has become an increasingly important asset of an organization. If the quality of the database falls short, even the best data mining techniques will result in poor predictions. This dissertation aims to bring traditional CRM models to a higher level by creatively incorporating new types of variables into these models. Typically a company’s database contains only some socio-demographic variables and data about the interactions between company and customers. However, to achieve a more complete view of the relation between customer and company it is desirable to augment these traditional datasets with new information types. These variables can be creatively collected and incorporated by the company itself or also attracted from external sources such as external data vendors, governmental agencies and weather institutes for example.

Data augmentation type	Authors	Stage of customer lifecycle
Web usage data	Hu & Zhong (2008)	Development
	Van den Poel & Buckinx (2005)	Development
Email interactions	Coussement & Van den Poel (2009)	Retention
Network-based variables	Hill, Provost & Volinsky (2006)	Acquisition
	Benoit & Van den Poel (2012)	Retention
Surveyed variables	Lix, Berger & Magliozzi (1995)	Acquisition
	Buckinx, Verstraeten & Van den Poel (2007)	NA
Geographical variables	Steenburgh, Ainslie & Engebretson (2003)	Acquisition
	Yang & Allenby (2003)	Acquisition

Table 1: Overview of past research about data augmentation

Although most research about aCRM models is focused on the data mining part, several authors have already proven that also data augmentation can be valuable for CRM models. In table 1 an overview is given of the different types of data augmentation, the authors who performed research in this field and the stage of the customer lifecycle on which this study focused. Both the studies of Hu & Zhong (2008) and Van den Poel & Buckinx (2005) suggest that combining traditional transactional and demographic data with web usage data can significantly improve purchasing behavior analysis. In comparison to traditional retailers who are only able to capture information of the final purchasing behavior of each client, online data can provide much more information. Online, a company can also keep track of how customers interact with the website during the whole shopping process. Even this information can be collected for website visits that eventually do not result in a purchase, which gives a more complete insight into the customers' purchasing behavior.

Other research by Coussement & Van den Poel (2009) describes how email interactions between the client and the company can assist a traditional churn prediction model. This study uses a computerized text analysis program to process huge amount of textual information from call center emails and translate this into positive and negative emotionality indicators. The incorporation of these indicators next to traditional RFM variables resulted in significantly better churn predictions.

Recently, several studies have incorporated network-based variables into CRM models. For example, based on telecommunication data, Hill, Provost, & Volinsky (2006) provides evidence that whether and how well a customer is linked to existing customers is a powerful characteristic to predict product adoption. Also Benoit & Van den Poel (2012) proved the added value of kinship network information to improve customer retention in financial services. These network-based marketing studies point out that traditional CRM models assume that customers act independently. However in reality, a customer's behavior is often influenced by friends, neighbors, family, other customers, etc., which should be incorporated in the models.

An important disadvantage of transactional data, typically used in database marketing, is that it only gives information about customers' behaviors without knowledge about the underlying reason for this behavior. In order to solve this issue, this behavioral data can be combined with surveyed attitudinal data. An important limitation of surveyed data though is that it typically is collected for only a limited number of respondents. Hence, Lix, Berger, & Magliozzi (1995) described how predictive models can contribute in linking this limited surveyed data with large commercially available databases. Also Buckinx, Verstraeten, & Van den Poel (2007) demonstrated how data mining techniques can be applied to make customer's loyalty predictions based on a small number of customer surveys. These predicted attitudes could then be used to enrich traditional transactional databases and improve other CRM models. However, this was not included in the paper.

A last example of data augmentation is the incorporation of geographical information of the customers. Although geographical data has already been used for a long time in customer segmentation (Baier, 1967), only recently, as a result of the increase in computational power, this variable can also be effectively included in predictive CRM models. In the study of Steenburgh, Ainslie, & Engebretson (2003) a hierarchical model is implemented to incorporate a highly categorical variable, such as zip codes. This results in a significantly better identification of potential customers for a private university. Also Yang & Allenby (2003) used an autoregressive approach to incorporate customer interdependence based on geographic proximity and democratic proximity. In this study, geographic reference groups turned out to be more important than demographic reference groups in predicting a customer's preference for Japanese cars. Similar as in network-based marketing studies, these studies try to take into account that consumer behavior can also be influenced by other neighboring customers.

4. Dissertation structure

This dissertation consists of research papers structured in chapters in such a way that each chapter can also be read independently. Table 1 gives an overview of the four studies in this dissertation in combination with information about the stage of the customer lifecycle in which this study has been applied, the data mining techniques used and the manner in which a traditional CRM model is improved by creatively augmenting the database with extra information. Most of the research papers discussed in this dissertation are situated in the CRM discipline of customer acquisition. Compared to customer development and customer retention models, customer acquisition models suffer the most from a lack of data quality. Since company databases typically contain only information about their own customers, the information used for customer acquisition is mostly limited to socio-demographic variables and, in the best case, added with lifestyle variables obtained from an external data vendor. Particularly in such a situation, creatively including other data types can significantly improve the predictive performance of a CRM model. However, the study discussed in Chapter III proves that also on top of transactional RFM variables used in a customer development model, data augmentation can be very valuable. In the fourth column of Table 1, the applied data mining techniques are presented for each chapter. Basically, each study consists of a basic traditional model and an augmented model. In order to improve the comparability, each basic model is built based on a logistic regression model. However, in order to create or incorporate new types of variables, often other, less traditional data mining techniques will be applied. Further, the last column in Table 1 shows that three types of data augmentation variables will be discussed in this dissertation: surveyed spending pleasure variables (Chapter II), situational variables (Chapter III) and spatial variables (Chapter IV & V).

Chapter II is related to the study of Lix et al. (1995), previously discussed, in which surveyed data of a limited number of respondents is linked with a commercially available database using predictive data mining techniques. Though, in comparison to the study of Lix et al. (1995), Chapter II applies a more modern data mining technique to extrapolate the surveyed constructs. Since the extrapolation of surveyed data is typically based on a limited number of observations compared to the number of independent variables, overfitting is likely to occur. Therefore this study uses random forests, a data mining technique proposed by Breiman (2001) that grows an ensemble of decision trees by combining random feature selection and bootstrap sampling techniques. This methodology has been argued to have excellent properties to avoid this problem of overfitting. Further, where Lix et al. (1995) mainly focused on comparing the predictive performance of the different extrapolation methodologies, the focus of this research is on the added value of the extrapolated variables itself. The study proposes a methodology for external data vendors to create commercial variables, called spending pleasure variables, based on surveyed purchasing behavior and attitudinal data. This methodology is applied in 26 product categories, creating spending pleasure predictions for more than 3 million respondents. Moreover, in contrast to Lix et al. (1995), the added value of these variables are evaluated in a customer acquisition model for a magazine publisher.

Chapter	Title	Stage of customer lifecycle	Data mining techniques	Data augmentation
II	Data Augmentation by Predicting Spending Pleasure using Commercially Available External Data [1]	Acquisition	- Logistic regression - Random forests	Surveyed purchasing behavior and attitudinal data
III	Improving Purchasing Behavior Predictions by Data Augmentation with Situational Variables [2]	Development	- Logistic regression - Generalized linear mixed model	Situational variables: - Physical surroundings - Temporal perspective - Social surroundings
IV	Improving Customer Acquisition Models by Incorporating Spatial Autocorrelation at Different Levels of Granularity [3]	Acquisition	- Logistic regression - Autologistic model	Spatial Interdependence
V	Including Spatial Interdependence in Customer Acquisition Models: a Cross-Category Comparison [4]	Acquisition	- Logistic regression - Autologistic model - Generalized linear mixed model	Spatial Interdependence

[1] Published in Journal of Intelligent Information Systems (2011)
[2] Published in International Journal of Information Knowledge and Decision Making (2010)
[3] Under review at Journal of Intelligent Information Systems
[4] Published in Expert Systems with Applications (2012)

Table 2: Overview of research papers

Chapter III points out that the purchasing behavior of a customer do not only depend on the characteristics of the individual, but can also be influenced by situational factors during the purchase occasion itself. The existence of situational influences during a customer's purchasing behavior has already been proven by (Belk, 1975), but this dissertation will investigate in a predictive CRM context how situational variables can add extra predictive value on top of traditional RFM variables. Three dimensions of situational variables will be examined: physical surroundings, temporal perspective and social surroundings respectively represented by weather, time and salesperson variables.

Chapter IV is positioned in the field of data augmentation with spatial information. This study is related to the research of Yang & Allenby (2003) that uses an autoregressive approach to take the interdependence between neighboring customers into account. Similar to Yang & Allenby (2003), also in Chapter IV an autologistic regression model is applied to incorporate these effects. In addition, this study investigates how the chosen granularity level on which these spatial effects are measured can have an effect on the predictive improvement of the model. Since these effects can have several origins (i.e. social influence, homophily and exogenous shocks), this study indicates that the autocorrelation between customers' behavior can be split into several parts, each optimally measured at different levels of granularity. Therefore, a model is introduced that simultaneously incorporates multiple levels of granularity in order to improve a customer identification model of a Japanese car brand. Further, also the effect of the sample size on these models is examined in detail.

Chapter V also investigates the incorporation of spatial interdependence in CRM models. Firstly, methodologically, this study compares the predictive performance of two models that are able to include spatial interdependence based on a geographic variable that groups customers into mutually exclusive neighborhoods. Besides an autologistic regression model as used in the research of Yang & Allenby (2003), Steenburgh et al. (2003) demonstrated that also a generalized linear mixed model is able to incorporate this effect. Secondly, besides comparing the predictive performance of two models, this study also examines the added value of incorporating spatial interdependence over several product categories, such as publicly consumed durable goods, privately consumed durable goods and even consumer packaged goods.

Finally, Chapter VI provides a general summary about the main findings of the different studies in this dissertation and discusses limitations and directions for future research.

5. References

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., & Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*, *156*(2), 508-523.

Baier, M. (1967). Zip Code - New Tool for Marketers. *Harverd Busness Review*, *45*(1), 136-140.

Belk, R. W. (1975). Situational Variables and Customer Behavior. *Journal of Consumer Research*, *2*(3), 157-164.

Benoit, D. F., & Van den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications*, (Forthcoming).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32.

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection☆. *Expert Systems with Applications*, *26*(4), 509-518.

Buckinx, W., Verstraeten, G., & Van den Poel, D. (2007). Predicting customer loyalty using the internal transactional database. *Expert Systems with Applications*, *32*(1), 125-134.

Bult, J. R., & Wansbeek, T. (1995). Optimal Selection for Direct Mail. *Marketing Science*, *14*(4), 378-394.

Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, *40*(2), 339-354.

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, *36*(3), 6127-6134.

De Bock, K. W., & Van den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, *38*(10), 12293-12301.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2012). Valuing customers. *Journal of Marketing Research*, *41*(1), 7-18.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science*, *21*(2), 256-276.

Hu, J., & Zhong, N. (2008). Web Farming With Clickstream. *International Journal of Information Technology & Decision Making*, *7*(2), 291-308.

Hung, C., & Tsai, C.-F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert Systems with Applications*, *34*(1), 780-787.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., et al. (2005). Choice Models and Customer Relationship Management. *Marketing Letters*, *16*(3-4), 279-291.

Lix, T. S., Berger, P. D., & Magliozzi, T. L. (1995). New customer acquisition: prospecting models and the use of commercially available external data. *Journal of Direct Marketing, 9*(4), 8-18.

Mccarty, J., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656-662.

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602.

Petrison, L. A., Blattberg, R. C., & Wang, P. (1993). Database Marketing: Past, Present, and Future. *Journal of Direct Marketing*, *7*(3), 27-43.

Prinzie, A., & Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems*, *44*(1), 28-45.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: quality comes to services. *Harvard business review*, *68*(5), 105-111.

Shin, H., & Cho, S. (2006). Response modeling with support vector machines. *Expert Systems with Applications*, *30*(4), 746-760.

Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively Categorical Variables□: Revealing the Inforraation in Zip Codes. *Marketing Science*, *22*(1), 40-57.

Teo, T. S. H., Devadoss, P., & Pan, S. L. (2006). Towards a holistic perspective of customer relationship management (CRM) implementation: A case study of the Housing and Development Board, Singapore. *Decision Support Systems*, *42*(3), 1613-1627.

Thomas, S., Blattberg, R. C., & Fox, E. J. (2004). Recapturing Lost Customers. *Journal of Marketing Research*, *41*(1), 31-45.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, *166*(2), 557-575.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, *157*(1), 196-217.

Yang, S., & Allenby, G. M. (2003). Modeling Interdependent Preferences. *Journal of Marketing Research*, *40*(3), 282-294.

Zahavi, J., & Levin, N. (1997). Applying Neural Computing to Target Marketing. *Journal of Direct Marketing*, *11*(1), 5-22.

CHAPTER II:

DATA AUGMENTATION BY PREDICTING SPENDING PLEASURE USING COMMERCIALLY AVAILABLE EXTERNAL DATA

This chapter is based upon Baecke, P. & Van den Poel, D. (2011). Data Augmentation by Predicting Spending Pleasure using Commercially Available External Data, *Journal of Intelligent Information Systems, 36*(3), 367-383.

CHAPTER II:
DATA AUGMENTATION BY PREDICTING SPENDING PLEASURE USING COMMERCIALLY AVAILABLE EXTERNAL DATA

Abstract

Since customer relationship management (CRM) plays an increasingly important role in a company's marketing strategy, the database of the company can be considered as a valuable asset to compete with others. Consequently, companies constantly try to augment their database through data collection themselves, as well as through the acquisition of commercially available external data. Until now, little research has been done on the usefulness of these commercially available external databases for CRM. This study will present a methodology for such external data vendors based on random forests predictive modeling techniques to create commercial variables that solve the shortcomings of a classic transactional database. Eventually, we predicted spending pleasure variables, a composite measure of purchasing behavior and attitude, in 26 product categories for more than 3 million respondents. Enhancing a company's transactional database with these variables can significantly improve the predictive performance of existing CRM models. This has been demonstrated in a case study with a magazine publisher for which prospects needed to be identified for new customer acquisition.

1. Introduction

Among business practitioners and marketing scientists today, there has been a shift in focus from the traditional mass marketing to customer relationship management (CRM) (Kannan and Rao, 2001). This is reflected by the expanding number of articles on CRM that have recently been published in the literature (Kamakura et. al., 2005). Earlier, one-to-one marketing was laborious, time-consuming and costly, but in recent years the rise of new media such as the internet enabled companies and their customers to communicate in a more direct manner and exchange information valuable to each other (Van den Poel and Buckinx, 2005). Moreover, the significant drop in costs of data warehousing and the exponential increase in computational power contributed to the fact that plenty of organizations started to acquire transactional data of their clients (Bult and Wansbeek, 1995; Petrison et al., 1993). Consequently, customer databases of huge magnitude are created and processed in order to get more insights into their consumers' buying behavior which should help to improve the marketing strategies.

This process of collecting and analyzing a firm's information regarding customer interaction in order to enhance the customers' value to the firm has been studied extensively in the marketing literature (Kamakura et. al., 2005). Analytical CRM can be used in a variety of stages of the customer lifecycle. Most research has been done on customer churn, which is focused on detecting those customers who have a high probability of leaving the company. This should enable the company to make the correct interventions in order to increase loyalty and prolong the lifetime of a customer. Customer retention has received a lot of attention in the domain ever since it has proven that even a small reduction in customer defection can have a great impact on a firm's profitability (Reichheld and Sasser, 1990; Van den Poel and Larivière, 2004; Gupta et al., 2004). The value of a customer can also be enhanced through customer development activities such as cross-selling and up-selling. Cross-selling is involved with encouraging customers to buy across categories, while up-selling is focused on increasing the demand of customers in existing categories (Prinzie and Van den Poel, 2006 and 2008; Ansari et al., 2000). Besides the immediate profit, both techniques deepen the customer relationship by increasing the share of products that is purchased at the company, thereby increasing the switching costs associated with purchasing from a competitor. Before a company is able to enhance their customer relationship, they first need to attract these customers. Customer acquisition is another stage of the customer life cycle where CRM can contribute useful insights. The objective in this domain is to attract more profitable customers.

In recent years, academic work in the field of direct marketing has focused on the development, improvement and comparison of new statistical techniques. These techniques are mainly used for segmentation or response modeling. Market segmentation involves dividing the total market into different clusters that are internally homogenous and mutually heterogeneous (Hung and Tsai, 2008). The desires of each cluster should be responded to with separate marketing actions. Response modeling refers to the use of costumer information in order to predict whether a customer will reply to a certain marketing action. Marketers will send mails or catalogs only to those consumers who have a high response probability and spend a large amount of money (Suh et al., 1999). A well-targeted mail increases profit while an irrelevant mail not only increases marketing cost but can also affect the customer-company- relationship in a negative way (Kim et al., 2008). Over the recent years, database marketing techniques have evolved from RFM models (based on the recency, frequency and monetary value of customer purchases) to statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression (Bult and Wansbeek, 1995; McCarty and Hastak, 2007). Recently, more advanced machine learning techniques were introduced like support vector machines, neural networks and random forests (Shin and Cho, 2006; Zahavi and Levin, 1997).

Besides the data mining technique that has been used, also the precision and depth of the database will have an important influence on the performance of such a response model and the potential of data analyses to increase profitability. The customer database can be seen as the foundation of CRM which will be used as input for the data mining techniques. The omission of relevant variables can lead to incorrect interpretations and poor predictions. In other words, if the quality of the data is inferior, even the best data mining techniques will still result in mediocre performance (Petrison et al., 1993; Verhoef et al., 2003). As a result, companies constantly try to augment their database through data collection as well as trough the acquisition of commercially available external data.

The remainder of this paper is organized as follows: In Section 2 the limitations of a classic transactional databases are discussed, and commercially available external data is presented as a solution. Section 3 presents the purchasing behavior and attitude matrix on which the creation of spending pleasure variables is based. The complete methodology to create these variables is elaborated on in Section 4. Section 5 demonstrates the value of these variables in a case study with a magazine publisher. Finally, conclusions and directions for further research are given in Section 6.

2. External databases as a solution for database limitations

Although companies try to improve their database quality by collecting data themselves, these transactional databases will still suffer from a couple of limitations. First of all, these databases are typically single source in nature. The data collection is limited to the information a company retrieves from their own customers which often results in an inward-looking view of the customer, as competitive information is mostly impossible to obtain. These databases do not capture the purchasing behavior of its customers in the total product category. Hence, the company has no indication about the total potential of each customer (i.e., the total needs of the customer for products in a certain product category) (Buckinx et al., 2007). However, this information can be extremely valuable in several applications. For example, when a company would like to target existing customers in a cross- or up-selling case, this information can help direct marketers to focus their marketing action on clients with a rather low purchasing behavior at the company in proportion to their full potential in the product category. The issue of single source data is even a bigger problem when a company wants to attract new customers. Because these persons have never had any contact with the company, no information is available to efficiently target prospects. Secondly, a great part of the data collected by most companies focuses on the past behavior of the individual. Although some authors recognize the predictive value of transaction information summarized in variables such as recency, frequency and monetary value, others remark that relational information should not be ignored and provide important additional insights to the company (McCarty and Hastak, 2007). Focusing only on transactional information is a very sales-oriented approach without understanding the underlying attitudes and motivation of the customer. Such an emphasis may increase sales in the short run, but does not improve the long-term relationship with the customer (Zahay et al., 2004). For example, when a company wants to acquire new customers, it would be easier to attract the customers who are not committed to the product category. This would increase sales at short notice. However, it will be very difficult to build a long relationship with such customers. Hence, in the long run, it would be more profitable to target customers who have a positive attitude towards the product category.

In order to solve these two limitations companies can enhance their databases with commercially available databases sold by external data vendors (Lix et al, 1995). Such databases differ from traditional company databases in two ways. First of all, they contain a large amount of demographic, socio-economic and life style variables for an extensive population. Such non-behavioral information can only be derived by directly questioning the respondent. Due to financial reasons and because it is impossible to reach every respondent in the database at all time, data about every variable is only available for a limited number of respondents in the total database of the external data vendor. A second characteristic of commercially available external databases is the fact that they are not related directly to a specific brand and often not to a specific product category (Lix et al, 1995). In other words, when a particular company wants to enhance his database with commercially available external data, it often has to deal with a lot of irrelevant variables and a large amount of missing values that are difficult to interpret. This is probably an important reason why the majority of the companies do not buy external data about their current customers (Verhoef et al., 2003).

To the best of our knowledge, little research has been done on the usefulness of commercially available external data for customer relationship management. Only Lix et al. (1995) described these databases and explored the linking of them to limited individual based survey data. But in this study they did not focus on the usefulness of the linked variables themselves. Moreover, they did not test the extra value of these variables when used for database enhancement of a company. In this study we will present a methodology that should help external data vendors to create variables that are a solution to the limitations stated earlier. These variables, called spending pleasure variables, have the following advantages. First of all, the variables are created based on a combination of behavioral information and attitudinal information surveyed for a limited number of respondents. The addition of relationship information over behavioral information will help the company to identify customers who are accessible for a long term relationship (Zahay et al., 2004). Secondly, due to financial reasons it is infeasible to obtain this information for each individual respondent in the external database. Consequently, in this study, we will show that it is sufficient to collect this information for only a limited number of respondents and use a predictive data mining technique (i.e. random forests) to extrapolate this spending pleasure information to all respondents in the commercially available external database in a similar way as in the study of Buckinx et al. (2007) and Lix et. al. (1995). As a result predictions for a large amount of respondents were created, which will make the variables more useful in a CRM context. Enhancing a company's single source database with this information will improve the performance of existing CRM models which will have a positive effect on the profitability. Moreover, because this data is also available for customers who are not a client of the company yet, it can also be used to identify prospects for acquisition. Thirdly, the respondents were questioned directly about specific product categories. This variable will be predicted based on a large amount of variables who are not relevant on their own for a direct marketer, but combined they can help to predict the spending pleasure variable, which is much more interpretable by managers.

3. The purchasing behavior and attitude matrix

Previous studies have indicated that transactional information, like RFM variables for example, are very valuable in predicting response and can increase profits in short term, but it would be advisable to also take relational information into account. Although this data have less predictive ability, it may be enormously useful in understanding the underlying tendencies and identifying those customers who are approachable for a long-term relationship (McCarty and Hastak, 2007; Zahay et al., 2004). As a result, we will predict spending pleasure variables based on a two-dimensional matrix including purchasing behavior and attitude in a product category. This matrix, displayed in Figure 1, is constructed in a similar way as in the study of Bandyopadhyay and Martell (2007), where loyalty was split into a behavioral and an attitudinal dimension.

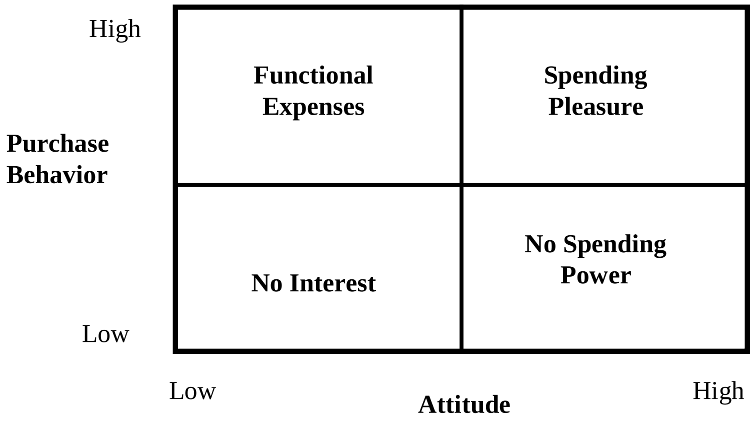


Figure 1: The purchasing behavior and attitude matrix

Based on this matrix we can position every surveyed customer in one of the following four quadrants. There is little that can be done with a person who is a non-user and is characterized with a weak attitude towards a certain product category. This person has no interest at all in the category and little resources should be wasted in an attempt to convince these individuals to buy products from this category. A respondent who spends a lot of money in a product category but doesn't have any affection with these products will be classified in the functional expenses quadrant. On the other hand, a respondent with a low purchasing behavior and a high attitude could potentially become a person with a lot of spending pleasure in the product category, but due to financial reasons his spending power is limited (Rossiter, 1995). This study aims to identify the respondents in the spending pleasure quadrant. These are the most valuable customers in the product category because they are big spenders and committed to the product category.

4. Methodology

4.1. Data description

This study will be based on data of one of the largest external data vendors in Belgium. This database includes about 10000 socio-demographic, economic and life-style variables of more than 3 million respondents. A typical example of the information in the database can be found in Table 1. Such a database is compiled by linking the data of many large, mostly online, surveys. Of course, not every respondent in the database has responded on all surveys. As a result the database suffers from a high number of missing values.

Socio- demographic	Economic	Lifestyle
Age group	Vehicle information	Leisure activities
Number of household members	Newspaper and magazine subscriptions	Favourite radio and television station
Gender	Average telecom payments	Sports
Life stage	Bank accounts	Favourite products
Social class	House information	Cultural interests

Table 1: Example of commercially available external data

4.2. Data collection

Because no data about the purchasing behavior and attitude towards specific product categories was available in the commercially available external database, it was necessary to survey this information from a limited number of respondents. The creation of the purchasing behavior construct is similar to the construct used by Dahl et al. (2001). Table 2 represents an overview of the four items measured on a seven point semantic differential scale. This scale consists out of the three RFM items and one item measuring the familiarity with the product category. The attitude construct on the other hand is constructed based on five items displayed in Table 3, also measured on a seven-point semantic differential scale. The selection of these items was based on other studies which involved the measurement of attitude towards a certain product category (Voss et al, 2003; Martin et al., 2001).

In comparison with your friends, how often do you purchase (product category)?	
	never or very rare - very often
How familiar are you with the purchase of (product category)?	
	not experienced - very experienced
In comparison with your friends, when was the last time you purchased (product category)?	
	very recent - never or very long ago (reversed)
In comparison with your friends, how much money do you spend at (product category)?	
	no or little money - a lot of money

Table 2: The purchasing behavior items

I consider (product category) to be:	
	unpleasant – pleasant
	exciting – dull (reversed)
	awful - delightful
	enjoyable – unenjoyable (reversed)
	boring – fun

Table 3: The attitude items

active sports	decoration	multimedia equipment	personal hygiene
cars	extra insurance	newspapers	phoning
cell phone	faster internet	non-profit	risk investments
cleaning products	food and drinks	no-risk investments	holidays
clothes	grocery	omnium insurance	wellness
consumer credit	magazines	passive sports	
culture	multimedia	pay-tv	

Table 4: Product categories overview

4.3. Survey response

In this study spending pleasure variables will be created for 26 product categories. An overview of these product categories can be found in Table 4. Because responsiveness to lengthy questionnaires has decreased and it would be too repetitive to question one respondent for all 26 product categories, the questionnaires were compiled so that the number of surveyed categories was limited to a maximum of five or six per respondent, similar to the one discussed in the study of Kamakura and Wedel (2003).

	Cronbach's Alpha	
Product category	**Purchasing behavior**	**Attitude**
active sports	0.9146	0.9368
cars	0.8202	0.8681
cell phone	0.8192	0.8914
cleaning products	0.8591	0.9374
clothes	0.8086	0.8988
consumer credit	0.8664	0.8993
culture	0.9296	0.9619
decoration	0.8530	0.9424
extra insurance	0.8814	0.9284
faster internet	0.8534	0.9079
food and drinks	0.7497	0.8976
grocery	0.7367	0.9003
magazines	0.9003	0.9151
multimedia	0.8999	0.9236
multimedia equipment	0.8758	0.9257
newspapers	0.9386	0.9447
non-profit	0.9362	0.9463
no-risk investments	0.9395	0.9491
omnium insurance	0.9212	0.9109
passive sports	0.9082	0.9612
pay-tv	0.9282	0.9497
personal hygiene	0.8831	0.9288
phoning	0.8270	0.9420
risk investments	0.9351	0.9303
holidays	0.8393	0.8693
wellness	0.9535	0.9679

Table 5: Cronbach’s Alphas per product category

150000 respondents or about 5% of the total external database were randomly addressed with an online questionnaire about their purchasing behavior and attitude in several product categories. 22083 persons responded on questions about at least one product category which results in a response rate of 14.72%. After elimination of bad and inconsistent respondents we maintained on average 3178 respondents per product category.We tested construct reliability per product category by means of Cronbach's coefficient alpha. All coefficients, presented in Table 5, clearly exceed the 0.7 level recommended by Nunnally and Bernstein (1994), which proves we use reliable constructs, especially given the fact that reversed coding was used to measure certain items.

4.4. Identification of spending pleasure respondents

Based on the construct scores we can position every surveyed respondent in the purchasing behavior and attitude matrix. By means of a cluster analysis these respondents can be divided into four segments that correspond with the four quadrants discussed earlier in this study. Respondents who are member of the segment with a high purchasing behavior and a high attitude towards a certain product category will be classified as having spending pleasure for that product category. This dummy variable will be the dependent variable in the predictive models used to make spending pleasure predictions for all members of the commercially available external database.

4.5. Prediction of spending pleasure using random forests

For the extrapolation of the spending pleasure variables over the total database we opted for the use of random forests, a machine learning technique introduced by Breiman in 2001 based on the principle of a decision tree (Breiman, 2001). A decision tree can be described as a flow of decision rules and their outcomes. Each node corresponds to a variable and a leaf represents a possible value of the target variable. It classifies an example by starting at the root of the tree and moving through the decision nodes until a leaf is reached, which provides the classification of the instance. This is a very popular technique because of its ease and interpretability, which is especially useful in a business context (Duda et al., 2001). Moreover, the technique is flexible in terms of input features and able to handle covariates at different measurement levels. A major drawback of this technique is its instability or lack of robustness (Hastie, Tibshirani and

Friedman, 2001). Small variations in data structure or feature space often result in very different series of splits, tree structures and predictions.

A solution for this problem is given by Random Forests. This algorithm combines a large number of decision trees that are constructed based on an independently sampled random vector with the same distribution for all trees in the forests. As a result of the Strong Law of Large Numbers the generalisation error converges to a limit without overfitting. The independently sampled random vectors are created by combining random feature selection and bootsrap sampling techniques. Breiman (2001) illustrated that the accuracy of Random Forests depends on the the strength of the individual classifiers but also the dependence between them. As a result, random feature selection is introduced to determine the split at each node. This injection of randomness should minimize the correlation between the classifiers and improve accuracy. In tandem with random feature selection, bagging or bootstrap sampling techniques are also included to improve accuracy. For each decision tree a new training set is drawn by sampling observations, with replacement, from the original training set. This should reduce variance and helps to avoid overfitting. Each tree is grown on a separate bootsrap sample and randomly selected features. Eventually, the forest will choose the classification having the most votes over all the trees in the forest.

Random forests have a number of advantages that are particularly attractive in this study, using a commercially available external database. First of all, this algorithm often outperforms classic predictive techniques like logistic regression (Coussement and Van den Poel, 2008). Secondly, as stated earlier external databases typically contain a lot of variables that have no or little predictive value. This technique has proven that the outcomes of the classifier are very robust when the data contains a lot of noise (Breiman, 2001). Thirdly, in this study we will build a model based on a limited number of respondents including a large number of variables. This can easily lead to over-fitting of the model, but because random forests is based on a large number of subset trees this problem is avoided. Finally, random forests are easy to implement because it includes a good method for estimating missing data and maintains accuracy when a large proportion of the data is missing. Further, there are only two free parameters to set. Based on the suggestions of Breiman (2001) the number of randomly chosen predictors was set equal to the square root of the total number of variables included in the model and 1000 trees were grown per random forests model.

Taking into account that a separate model for each of the 26 product categories has to be created and scored for more than 3 million respondents in the commercially available external database, it is computationally too expensive to include all 10000 variables into the random forests model. Although random forests have a random feature selection embedded in the algorithm and can deal with large feature sets, this database is still too big to work efficiently on. Therefore, preceding the random forests model, a simple maximum-relevance variable selection based on the correlation between the variables and the classification variable is performed to reduce the total number of input variables to 300 per product category, which is computationally more feasible to work with.

4.6. Predictive performance of the resulting models

In order to be able to evaluate the predictive performance of each of the 26 models, each surveyed sample was split into two parts. Firstly, the predictive model is estimated on a training set, containing 70% of the surveyed sample. Afterwards, this model is validated on the remaining 30% of the surveyed sample. It is essential to evaluate the performance of the classifiers on a holdout validation sample in order to ensure that there is no overfitting in the training model. The area under the receiver operating characteristic curve (AUC) is used as evaluation metric of the classifiers (Hanley and McNeil, 1982). The receiver operating characteristic (ROC) curve is a graphical plot of the sensitivity (i.e. the number of true positives versus the total number of events) and 1-specificity (i.e. the number of true negatives versus the total number of non-events) for all possible cut-off values used. The AUC measures the area under this curve and can range from 0.5, if the predictions are as good as random, to 1, if the model's predictions are perfect. The advantage of an AUC in comparison with other evaluation metrics, like the percent correctly classified (PCC), is the fact that PCC is highly dependent on the chosen threshold. The PCC gives only an indication of the performance at one cut-off, while AUC is a performance metric including all cut-off levels.

Product category	**AUC values**
non-profit	0.8193
active sports	0.8177
risk investments	0.8149
newspapers	0.8121
passive sports	0.7933
culture	0.7784
wellness	0.7559
pay-tv	0.7455
phoning	0.7437
multimedia equipment	0.7327
clothes	0.7284
consumer credit	0.7283
holidays	0.7230
omnium insurance	0.7194
cars	0.7095
faster internet	0.7073
multimedia	0.7029
no-risk investments	0.6968
cleaning products	0.6958
magazines	0.6956
cell phone	0.6951
decoration	0.6837
personal hygiene	0.6751
extra insurance	0.6669
food and drinks	0.6458
grocery	0.6378
Average AUC	**0.7279**

Table 6: Predictive performance (in terms of AUC) per product category

Table 6 ranks for each product category the AUC values of the models on the validation sample. The predictive performance varies from 0.6378 to 0.8193 with an average AUC of 0.7279. Apparently the commercially available external database includes more valuable data to identify spending pleasure respondents in product categories as non-profit, active sports, risk investments and newspapers than to predicting spending pleasure in product categories as grocery, food and drinks, extra insurance and personal hygiene.

Based on these models all respondents in the commercially available database were scored 26 times, once per product category. This results in the creation of 26 variables for 3,218,759 respondents indicating their probability of having spending pleasure in a certain product category.

5. Application

These spending pleasure variables can be very attractive for other companies to enhance their database because they solve a couple of shortcomings. These are easy interpretable variables containing information about the purchasing behavior and attitude as well in the total product category, whereas the classic database of a firm is mostly limited to socio-demographic and single-source transactional information between the customer and company. Moreover, the spending pleasures variables are known for a large amount of respondents, also non-customers of the company, and do not include missing values. Enhancing a company’s database with such variables will result in a better predictive performance of existing CRM models and increase the profitability. Especially in the case of new customer acquisition, for which buying external data is most popular because companies do not possess data about prospects on their own, these variables can be very valuable (Verhoef et al., 2002). In this section we will use the spending pleasure variables to enhance the database of a monthly issued magazine in order to improve the selection of prospects for new customer acquisition.

5.1. Research question of the company

An application of the spending pleasure variables has been implemented in cooperation with a magazine publisher. This monthly issued magazine is positioned in the market as a magazine specially designed for elderly people and contains information about topics such as law and finance, healthiness, people and opinions, leisure time, lifestyle and multi media. The main target group of this magazine are persons older than fifty. Consequently and not surprisingly, age is an important variable in identifying prospects.

In this study the commercially available external data is used in order to identify prospects with the same profile as the existing magazine subscribers. A logistic regression model will be built in order to predict the magazine subscribers. Based on this model all respondents in the external database will be scored and the company can target a top section of the respondents who are still not a client with the highest probability of being a subscriber. A comparison will be made between the predictive performance of the model based on data excluding and including the spending pleasure variables.

5.2. Methodology

Logistic regression with a stepwise feature selection was chosen in order to solve this binary classification problem because this is a statistical technique that is frequently used in the commercial world and has a better performance than other popular techniques as chi-square automatic interaction detection (CHAID) for example (Verhoef et al., 2002).

The analyses are performed on a database containing 125,434 respondents, consisting of 62,717 existing subscribers and the same amount of randomly chosen non-subscribers. The dependent variable will be the binary variable subscription (i.e. one for the subscribers and zero for the non subscribers). First, a model is built based on 125 socio-demographic independent variables. Subsequently, we enhance this database with 26 spending pleasure variables and compare the predictive performance. Both models will be built on a training sample of 70% of the total database and evaluated on a validation sample, containing the remaining 30% of respondents. AUC will be used to evaluate the predictive performance of both models.

5.3. Results

A model based on only socio-demographic data performs already very well with an AUC of 0.8045 on the validation sample. This was more or less expected since the magazine is positioned as a magazine for elderly people and the socio-demographic data contains several well discriminating age group variables. Despite the fact that this model performs already very well, which makes it more difficult to improve it, enhancing the data with only 26 spending pleasure variables lifts the AUC to a value of 0.8385 on the validation sample. This significant improvement in predictive performance of 0.0340 will result in better predictions of potential subscribers which will increase the success ratio of the acquisition campaign and improve the profitability.

All variables that have a significant influence (alpha = 0.05) on predicting magazine subscribers are presented in Appendix 1, ranked by the absolute values of their standardized betas. Looking at the top of this table it is clear that besides socio-demographic variables like age and gender, the spending pleasure variables contribute extra value to the database in order to improve the model's prediction. This table demonstrates that there is a negative relationship between all the age groups lower than fifty years old and the dependent subscription variable. This confirms the fact that this magazine is positioned as a magazine for elderly people. Also the spending pleasure variables are easily interpretable. Obviously, respondents with a high purchasing behavior and attitude toward magazines are more likely to subscribe to this magazine. But also variables as spending pleasure for holidays and omnium insurance have a positive relation with the magazine subscriptions. This is probably due to the topics leisure time as well as law and finance in the magazine.

6. Conclusion and directions for further research.

The emergence of customer relationship management in marketing resulted in the fact that the company's database becomes more and more important to improve customer relationships and attract new customers. Although companies are able to collect a large amount of data from their own customers, these transactional databases will still suffer from several limitations. Firstly, such databases contain only single source data coming from the company's own customers. They contain no information about non-customers or about the purchasing behavior of existing customers in the total product category. Secondly, these databases typically contain transactional information about the purchasing behavior of the customer, like recency, frequency and monetary value. Including attitudinal information could help to identify the customers who are committed to the product category and more approachable to build up a long term relationship with. These limitations can be solved by enhancing the company's database with commercially available external data. But among business practitioners, these external databases are not always very popular because they suffer also from a couple of drawbacks. Typically, these databases include a lot of missing values and most variables are not related directly to a specific brand or product category. Consequently, these variables are difficult to interpret and not attractive to enhance a company's database with. This study describes a methodology for an external data vendor to create variables that solve all of these limitations. The spending pleasure variables are composed of purchasing behavior and attitude dimension in specific product categories, predicted for a large amount of respondents (customers and non-customers) without missing values.

Such spending pleasure variables were created by questioning a limited number of respondents about their purchasing behavior and attitude in a specific product category. By combing these two constructs in a two dimensional matrix respondents in the spending pleasure segment can be identified. This dummy variable is predicted for all respondents in the commercially available database by means of the random forests predictive modeling technique. This results in the creation of 26 spending pleasure variables for more than 3 million respondents. These easily interpretable variables can be very valuable to a company and improve the predictive performance of existing CRM models. This has been demonstrated in a new customer acquisition case for a magazine publisher. Enhancing a predictive model based on socio-demographic variables with spending pleasure variables resulted in a significant increase of the AUC performance.

While we strongly believe that this research paper fills a large gap in today’s literature, there are still some directions for future research. Firstly, this study demonstrates the usefulness of spending pleasure variables in a new customer acquisition case. Future research can investigate how the spending pleasure variables perform in other contexts of the CRM field, like in cross-selling, up-selling or churn models. Secondly, in this study we predicted only the respondents who are positioned in the spending pleasure segment because these are valuable and interesting respondents for most companies, but in particular cases it could also be useful to identify the respondents who see the product category as a functional expense or have a lack of spending power. It is advisable to target these respondents with different communication strategies than the spending pleasure respondents. For example, customers with a lack of financial resources to spend in the product category but a high attitude towards the product category can still be convinced by offering easy credit facilities for the product.

Acknowledgement

We would like to thank the anonymous external data provider for supporting this research and for making the data available. We are also grateful to Leo Breiman for his inspiring research and the public availability of the Random Forests software.

Appendix 1: Significant socio-demographic and spending pleasure variables in the acquisition model

		Standard	Wald Chi		Standardized
Variable	Beta	Error	Square	P-value	Beta
age group 36-40	-4.1614	0.1161	1283.8251	0.0000	-0.5146
age group 41-45	-3.4665	0.0862	1618.6206	0.0000	-0.4740
age group 31-35	-4.4344	0.1432	959.1675	0.0000	-0.4594
age group 26-30	-4.6407	0.1966	556.9500	0.0000	-0.3922
age group 22-25	-5.2401	0.4508	135.1342	0.0000	-0.3076
age group 46-50	-1.9967	0.0629	1008.7171	0.0000	-0.2930
age group 18-21	-5.6122	0.9958	31.7604	0.0000	-0.2572
gender: female	0.7544	0.0342	487.4052	0.0000	0.2030
SP for magazines	2.0310	0.0969	439.2208	0.0000	0.1477
SP for holidays	1.1228	0.0584	369.1060	0.0000	0.1086
language: French	-0.3550	0.0340	108.8498	0.0000	-0.0943
number of household members	-0.1245	0.0130	91.1683	0.0000	-0.0840
SP for omnium insurance	1.2624	0.1380	83.6999	0.0000	0.0705
SP for cell phones	-1.5492	0.1353	131.1670	0.0000	-0.0673
SP for clothes	-0.7908	0.0895	78.1094	0.0000	-0.0637
% of unemployed people in the neighborhood	-0.0184	0.0021	76.0148	0.0000	-0.0603
% of higher educated people in the neighborhood	-0.0139	0.0022	39.2756	0.0000	-0.0510
SP for non-profit organizations	0.6718	0.0746	81.1742	0.0000	0.0500
SP for faster internet	0.6155	0.1042	34.9082	0.0000	0.0479
SP for phoning	-0.9283	0.1426	42.3819	0.0000	-0.0473
age group 51-55	-0.2401	0.0462	26.9907	0.0000	-0.0466
number of children between 12 and 15 in the household	-0.2528	0.0472	28.6546	0.0000	-0.0464
SP for consumer credit	1.1327	0.1632	48.2024	0.0000	0.0435
high status factor variable	0.0005	0.0001	25.5749	0.0000	0.0434
age group 71-75	0.2511	0.0324	60.1730	0.0000	0.0387
SP for no risk investments	0.6925	0.1104	39.3654	0.0000	0.0379
number of women between 51 and 55 in the household	0.1891	0.0393	23.1101	0.0000	0.0372
SP for risk investments	1.2698	0.2379	28.4932	0.0000	0.0364
SP for grocery	0.5012	0.1032	23.5987	0.0000	0.0362
number of children between 16 and 17 in the household	-0.2699	0.0535	25.4243	0.0000	-0.0348
age group 66-70	0.2044	0.0326	39.2156	0.0000	0.0343
SP for newspapers	0.4397	0.0764	33.1197	0.0000	0.0330
SP for active sports	-0.4646	0.0810	32.9254	0.0000	-0.0326
director of a private limited company	-0.3520	0.0604	33.9222	0.0000	-0.0313
number of women between 46 and 50 in the household	0.1962	0.0490	16.0042	0.0001	0.0313
presence of a phone	0.1060	0.0195	29.5658	0.0000	0.0292
number of men between 41 and 45 in the household	-0.1936	0.0620	9.7456	0.0018	-0.0268
age group 60-65	0.1399	0.0389	12.9077	0.0003	0.0266
SP for food and drinks	-0.4396	0.0948	21.4944	0.0000	-0.0265
number of men between 46 and 50 in the household	-0.1744	0.0489	12.7215	0.0004	-0.0255
number of children between 6 and 11 in the household	-0.1102	0.0510	4.6582	0.0309	-0.0244
number of men between 76 and 80 in the household	0.1947	0.0381	26.0608	0.0000	0.0235
number of men between 51 and 55 in the household	-0.1328	0.0353	14.1237	0.0002	-0.0233
number of women older than 80 in the household	-0.2053	0.0421	23.8074	0.0000	-0.0216

SP for personal hygiene	0.3253	0.1117	8.4746	0.0036	0.0204
number of men between 18 and 21 in the household	-0.1383	0.0425	10.5884	0.0011	-0.0193
neighborhood with Mediterranean people	-0.1776	0.0547	10.5652	0.0012	-0.0187
head of the household	0.0684	0.0272	6.3214	0.0119	0.0187
SP for extra insurance	0.2939	0.1024	8.2357	0.0041	0.0185
number of women between 61 and 65 in the household	0.0974	0.0362	7.2217	0.0072	0.0182
high carrier people	-0.2971	0.0822	13.0665	0.0003	-0.0182
SP for multimedia	-0.2848	0.0924	9.4949	0.0021	-0.0171
life stage: middle age	0.0002	0.0001	8.5214	0.0035	0.0157
older couples	0.1535	0.0468	10.7533	0.0010	0.0156
number of women between 18 and 21 in the household	-0.1134	0.0444	6.5363	0.0106	-0.0155
Italian roots	-0.1702	0.0685	6.1840	0.0129	-0.0146
urban residential neighborhood	-0.1018	0.0356	8.1651	0.0043	-0.0138
SP for pay TV	0.4235	0.1906	4.9368	0.0263	0.0132
revenue class 1	0.0720	0.0300	5.7672	0.0163	0.0130
semi-urban residential neighborhood	0.0715	0.0280	6.5298	0.0106	0.0123
number of men older than 80 in the household	0.1245	0.0471	6.9806	0.0082	0.0120
revenue class 3	-0.0552	0.0234	5.5467	0.0185	-0.0116
household with teenagers	-0.1068	0.0518	4.2472	0.0393	-0.0105
rural zone	0.0775	0.0382	4.1167	0.0425	0.0101

References

A. Ansari, S. Essegaier and R. Kohli, Internet Recommendation Systems, Journal of Marketing Research 37 (3) (2000), p. 363-375.

S. Bandyopadhyay and M. Martell, Does attitudinal loyalty influence behavioral loyalty? A theoretical and empirical study, Journal of Retailing and Consumer Services 14 (1) (2007), p. 35-44.

L. Breiman, Random Forests, Machine learning 45 (1) (2001), p. 5-32.

G.C. Bruner, P.J. Hensel and K.E. James, Marketing Scales Handbook 4: A Compilation of Multi-Item Measures for Consumer Behavior & Advertising, Ohio: Thomson/South Western (2005).

W. Buckinx, G. Verstraeten and D. Van den Poel, Predicting customer loyalty using the internal transactional database, Expert Systems with Applications 32 (1) (2007), p 125-134.

J.R. Bult, T. Wansbeek, Optimal selection for direct mail, Marketing Science 14 (4) (1995) 378-394.

K.Coussement and D. Van den Poel, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications 34 (1) (2008), p. 313 -327.

D.W. Dahl, R.V. Manchanda and J..J. Argo, Embarrassment in Customer Purchase: The roles of Social Presence and Purchase Familiarity, Journal of Consumer Research 28 (3) (2001), p.473-481.

R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, New York: Wiley (2001).

S. Gupta, D.R. Lehmann and J.A. Stuart, Valuing Customers, Journal of Marketing 41 (1) (2004), p. 7-19.

J.A. Hanley and B.J. McNeil, The meaning and use of area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982), p.29-36.

T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: Data mining, inference and prediction, New York: Springer-Verlag (2001)

C. Hung and C. Tsai, Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand, Expert Systems with Applications 34 (1) (2008), p. 780-787.

W. Kamakura, C.F. Mela, A Ansari, A. Bodapati, P. Fader, R. Iyengar, P. Naik, S. Neslin, B. Sun, P.C. Verhoef, M. Wedel and R. Wilcox, Choice Models and Customer Relationship Management, Marketing Letters 16 (3) (2005) p. 279-291.

W.A. Kamakura and M. Wedel, List augmentation with model based multiple imputation: a case study using a mixed-outcome factor model, Statistica Neerlandica 57 (1) (2003), p.46-57.

P.K. Kannan, H.R. Rao, Introduction to the special issue: decision support issues in customer relationship management, Decision Support Systems 32 (2) (2001), P. 83-84.

D. Kim, H. Lee and S. Cho, Response modeling with support vector regression, Expert Systems with Applications 34 (2) (2008), p. 1102-1108.

T.S. Lix, P.D. Berger, T.L. Magliozzi, New Customer Acquisition: Prospecting Models and the Use of Commercially Available External Data 9 (4) (1995), p.8-19.

I.M. Martin and D.W. Stewart, The Differential Impact of Goal Congruency on Attitudes, Intensions, and the Transfer of Brand Equity, Journal of Marketing Research 38 (4) (2001), p471-484.

J.A. McCarty, M. Hastak, Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, Journal of business research 60 (6) (2007) p. 656-662.

J.C. Nunnally and I.H. Bernstein, Psychometric theory, New York: McGraw-Holl (1994).

L.A. Petrison R.C. Blatteberg and P. Wang, Database marketing past, present and future, Journal of direct marketing, 7 (3) (1993) p. 27-43A. Prinzie and D. Van den Poel, Exploiting Randomness for Feature Selection in Multinomial Logit: A CRM Cross-Sell Application, Lecture Notes in Artificial Intelligence 4065 (2006), p.310-323.

A. Prinzie and D. Van den Poel, Random Forests for Multiclass classification: Random Multinomial Logit, *Expert Systems with Applications*, 34 (3) (2008)

F. F. Reichheld, and W.E. Jr. Sasser, Zero defections: quality comes to services. Harvard Business Review, 68(5) (1990), p105-112.

J.R. Rossiter, “Spending Power” and the Subjective Discretionary Income (SDI) Scale, Advances in consumer research 22 (1) (1995), p. 236-241.

H. Shin and S. Cho, Response modeling with support vector machines, Expert Systems with Applications 30 (4) (2006), p. 746-760.

E.H. Suh, K.C. Noh and C.K. Suh, Customer list segmentation using the combined response model, Expert Systems with Applications 17 (2) (1999), p. 89-97.

D. Van den Poel and W. Buckinx, Predicting online purchasing behaviour, European Journal of Operational Reasearch 166 (2) (2005), p. 557-575.

D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Reasearch 157 (1) (2004), p. 196-217.

P. C. Verhoef, P.N. Spring, J. C. Hoekstra and P.S.H. Leeflang, The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands, Decision Support Systems 34 (4) (2003), p. 471-481.

K.E. Voss, E.R. Spangenberg and B. Grohmann, Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude, Journal of Marketing Research 40 (3) (2003), p.310-320.

J. Zahavi and N. Levin, Applying Neural Computing to Target Marketing, Journal of direct marketing 11 (1) (1997), p.5-23.

D. Zahay, J. Peltier, D.E. Schulz and A. Griffen, The Role of Transactional versus Relational Data in IMC Programs: Bringeng Customer Data Together, Journal of advertising research 44 (1) (2004), p.3-18.

CHAPTER III:

IMPROVING PURCHASING BEHAVIOR PREDICTIONS BY DATA AUGMENTATION WITH SITUATIONAL VARIABLES

This chapter is based upon Baecke, P. & Van den Poel, D. (2010). Improving Purchasing Behavior Predictions by Data Augmentation with Situational Variables, *International Journal of Information Technology & Decision Making*, *9*(6), 883 – 872.

CHAPTER III:
IMPROVING PURCHASING BEHAVIOR PREDICTIONS BY DATA AUGMENTATION WITH SITUATIONAL VARIABLES

Abstract

Nowadays, an increasing number of information technology tools are implemented in order to support decision making about marketing strategies and improve customer relationship management (CRM). Consequently, an improvement in CRM can be obtained by enhancing the databases on which these information technology tools are based. This study shows that data augmentation with situational variables of the purchase occasion can significantly improve purchasing behavior predictions for a home vending company. Three dimensions of situational variables are examined: physical surroundings, temporal perspective and social surroundings respectively represented by weather, time and salesperson variables. The smallest, but still significant, increase in predictive performance was measured by enhancing the model with time variables. Besides the moment of the day, this study shows that the incorporation of weather variables, and more specifically sunshine, can also improve the accuracy of a CRM model. Finally, the best improvement in purchasing behavior predictions was obtained by taking the salesperson effect into account using a multilevel model.

1. Introduction

In an increasingly competitive business environment, a successful company must provide customized services in order to gain a competitive advantage.[1] As a result, many firms have implemented information technology tools to customize marketing strategies in order to build up a long-term relationship with their clients.[2] The technological development, and more specifically the rise of the internet, have extended the opportunities of a firm to interact with the customer.[3, 4] Moreover, the continuing decline in costs for information processing and data warehousing makes the collection of historical purchasing behavior information even more attractive.[5]

This evolution is also reflected in the growing body of empirical research about customer relationship management (CRM).[6, 7] Among academic researchers, there exists a strong sense that CRM can improve marketing strategies resulting in higher profits.[8, 9] In general, CRM can be split into an operational and analytical part.[10] While operational CRM is focused on the automation of business processes, this study can be situated in the domain of analytical CRM. In analytical CRM a firm tries to collect and analyze data regarding customer interactions in order to create a deeper understanding of their customers' behavior, identify the most profitable group of customers and improve their value to the firm across the various stages of the customer lifecycle.[6] First of all, CRM can be used to identify profitable customers that are most suitable for acquisition.[11] Next, direct marketing tools, such as direct mail and coupons, are used to attract these customers.[12] Once the customers are acquired the firm should focus on customer retention.[13] Customized marketing actions are implemented to increase satisfaction and loyalty in order to stretch out the customer's lifetime at the firm. Due to the fact that the individual profitability of a customer increases over time, even a small improvement in customer retention can have a great impact on the firm's total profitability.[14, 15] Finally, CRM can also be used in order to increase the individual value of existing clients, called customer development. Promoting more profitable (up-selling) or closely associated products (cross-selling) are activities typically used for this marketing strategy.[16]

Due to the constant increase in automation of business processes, customer databases of huge magnitude are created.[17] Data mining techniques are often used in analytical CRM to transform this large amount of unstructured data into useful, structured and valuable knowledge that can be used to support marketing decision making and forecast the effect of it[18]. Based on such data mining techniques, customers can be segmented into clusters with internally homogenous and mutually heterogeneous characteristics.[19] Besides segmentation, customers can also be ranked on their probability to behave in a certain way (e.g. buying a specific product or responding to a certain marketing campaign). With the help of these segmentation schemes and rankings a firm is able to approach only carefully selected customers, resulting in a higher success rate of their marketing campaigns.[20] Nowadays, CRM would be impossible without data mining. Consequently, researchers often try to improve CRM by enhancing the data mining techniques themselves. As a result, the data mining techniques used for CRM have gone through a major evolution. RFM models (i.e. recency, frequency and monetary value of customer purchases), but also classification techniques such as chi-square automatic interaction detection (CHAID) and regression models are already used in CRM for a long time.[21, 22] Recently, researchers try to outperform these primitive techniques by introducing more advanced machine learning algorithms, like support vector machines, neural networks and random forests.[23, 24] A last trend to improve predictions is by combining the outcome of several data mining techniques in an ensemble approach.[16, 25]

Besides focusing on the data mining techniques, researchers can also improve CRM models by enhancing the customer database used as input for the data mining techniques. Companies must consider their customer database as one of their most important assets in order to enable state of-the art CRM. Inferior database quality will automatically result in a “garbage in garbage out” effect. Even with the best data mining techniques, the predictive performance of the CRM model will always be poor if the customer database falls short.[11] Although traditional transactional variables will always result in good predictive performance,[26] some researchers already demonstrated that data augmentation with alternative variables may significantly improve CRM models. In Ref. 27 geographic data is incorporated (i.e. ZIP-codes) in a hierarchical model to improve direct marketing campaigns for the attraction of new students. Ref 3 and Ref 28 suggest combining clickstream information with traditional variables, such as historical purchasing behavior and demographics, in order to improve online-purchasing behavior predictions. Based on consumer networks formed using direct interaction between consumers, additional network attributes are created in Ref. 29 for each prospect. Taking this network information into account resulted in an increase of response rates for product/service adoption. In Ref. 30, a computerized text analysis program is used to compile positive and negative emotionality indicators from call center emails. They indicated that incorporating these emotions in an extended RFM model helps to better identify potential churners. One way to improve data quality and enhance a firm’s database is by purchasing commercially available data from an external data vendor.[31] Ref. 11 describes a methodology to create commercially available variables, a composite measure of purchasing behavior and attitude, that can be used for data augmentation and provide additional predictive performance to CRM models, especially in customer acquisition models.

The focus of this study will also be on data augmentation by investigating how situational variables are able to improve purchasing behavior predictions. Traditional CRM models are typically based on variables related to the individual (e.g. socio-demographics, individual past purchasing behavior). This study points out that the purchasing behavior of a particular customer can also depend on the situation of the purchase occasion itself. To the best of our knowledge, only a limited amount of past academic research recognized that situational variables can help to explain and understand consumer behavior,[32, 33] but none of these studies ever used situational variables for data augmentation to improve purchasing behavior predictions.

The remainder of this paper is organized as follows: Section 2 elaborates on situational variables and introduces three situational dimensions that will be incorporated in the model. The methodology is described in Section 3, consisting of the data description, the classification techniques used in this study and the evaluation criterion. Section 4 reports the empirical results. Finally, conclusions and directions for further research are given in Section 5.

2. Situational Variables

Generally, most CRM models are based on only individual variables such as socio-demographics, lifestyle variables and the individual past purchasing behavior of the customer. This study suggests that the situation in which the purchase occasion takes place can also play a significant role on the customer's choice. Although the amount of research specifically focused on situational influences is still small, a number of studies found evidence that situations can affect consumer behavior systematically.[32, 33] Despite these findings, situational variables never were used for data augmentation in a CRM context. This is mainly because predictions are usually made well before the purchase occasion takes place, which makes it difficult to take situational variables into account. But often, some of these variables are already known in advance. For example, in the home vending industry, the company decides when to visit which customer. This makes it possible to already include some situational characteristics in a highly dynamic model that scores the customers on a daily basis. In Ref. 32, five dimensions of situational variables are defined: physical surroundings, temporal perspective, social surroundings, task definition and antecedent states. The focus of this study will be on the first three dimensions because these situational variables can easily be included in a CRM model without a large increase in extra costs.

Physical surroundings are the most evident features of a situation. These features include all material surroundings, but also surrounding factors such as location, sounds, aromas, weather and lighting. This study is based on data of a home vending company specialized in frozen foods and ice cream. For this last product category, it can be expected that weather, in particular sunshine and temperature, is an important physical surrounding. Although little research exists about the influence of weather on consumer behavior, the influence on human behavior and business activities has been explored in several fields. In the field of psychology, weather is believed to influence people's mood. Ref. 34 examined the effect of six parameters on mood and found significant main effects of temperature, wind power and sunlight on negative affect. In the field of finance, some researches even demonstrated a significant relationship between the amount of sunshine and stock market purchasing.[35-37] Because modern short-term weather predictions are very accurate, these variables can easily be used to enhance the database on which CRM models are based. Besides the current weather, also weather history of the last seven and thirty days will be included in the model.

Temporal perspective is a situational dimension related to time. Ref. 38 examined the relationship between two situational variables (i.e. store environment and time pressure) and shopping behavior. They found evidence that the time available for shopping significantly affects the frequency of failure to make intended purchases, unplanned buying behavior, brand switching and the purchase volume. Practically, time pressure is difficult to measure and consequently not possible to include in a CRM model. Alternatively, this study will incorporate the moment of the day (i.e. morning, afternoon or evening) when a salesperson visits the client.

Social surroundings refer to other persons present during the purchase occasion, their characteristics, influences and interpersonal interactions. In a home vending environment the most important social surrounding is the interaction between the customer and the salesperson. A salesperson's personal attitudinal and behavioral characteristics have an important impact on his sales performance.[39] In this study we assume that purchase occasions of the same salesperson are correlated with each other. Hence a multilevel model is introduced to capture this effect.

The two other situational dimensions (i.e. task definition and antecedent states) will not be included in the model because they are related to specific motivations and attitudes of the customer. Task definition refers to the underlying motive why a customer will buy a particular product (e.g. a gift or personal use) and antecedent states include the momentary mood of a customer. This information is not available in a traditional transactional database and would be too costly to obtain for every customer. Hence, this study will only focus on the first three dimensions of situational variables that are practically implementable.

In a home vending environment, the visit schedule is mostly created at least one day in advance. Once this is finished, the decision maker already knows at what time and which salesperson will visit a particular customer. Besides this information, also weather predictions and historical weather information can be attracted without a lot of effort. In other words, in a dynamic CRM model that is scored on a daily basis, these three situational variables can easily be incorporated. This study will investigate whether data augmentation with such situational variables will result in better purchasing behavior prediction. These predictions generated daily can be used for several applications. For example, when the demand is too high to visit every client, these predictions can help to select the most profitable ones. On the other hand, in a situation of overcapacity, when the salesperson has extra time left, the predicted probabilities can be used to generate revisit suggestions of the most profitable clients that were not home.

3. Methodology

3.1. Data description

For this study, data is collected from a large home vending company, specialized in frozen foods and ice cream. This company uses about 180 salespeople to distribute their products to approximately 160,000 clients, visited on a regular basis in a biweekly schedule. Transactional data is used from February 1st, 2007 to November 30th, 2007 to build and validate the model. The same period in 2008 is used for out-of-period testing. Because a lot of promotional activities take place during the holiday period of Christmas and New Year, the months December and January are excluded and should be scored with a different model. For the creation of the weather variables, data about the daily sunshine and temperature has been obtained from the Belgian weather institute.

The data from the home vending company and the Belgian weather institute has been captured in explanatory variables. In Table 1, an overview of all variables used in this study can be found. The purpose of the proposed model is predicting whether a customer will buy at least one product conditional on him/her being at home. Therefore, only observations where the customer is at home are retained in the model. In a next step, this model can be combined with a second model predicting the probability a client will be at home, but this is beyond the scope of this research. In order to avoid correlation between purchase occasions of the same customer, only one visit per customer is randomly selected. If the customer was at home during the visit, (s)he bought at least one product in 46% of the purchase occasions. This signifies that the analysis table for this study is rather equally balanced between events and non-events.

Variable name	Description
Dependent variable:	
Sales	A binary variable indicating whether the customer purchased at least one product
Independent variables:	
Transactional variables:	
Recency visit	The number of days since the last visit
Recency bought	The number of days since the last purchase
Frequency visit	The number of visits in the last 8 weeks
Frequency bought	The number of purchases in the last 8 weeks
Monetary value	Total monetary value spent in the last 8 weeks
Sales ratio	The percentage of purchases based on all visits in the last 8 weeks
Avg. monetary value	The average amount spent per visit
Last time visit	A binary variable indicating whether the customer was visited in the last 21 days
Last time bought	A binary variable indicating whether the customer purchased at least one good at the last visit within 21 days
Last time amount	The amount spent on the last visit within 21 days
Weather variables:	
Sunshine	The total minutes of sunshine on the day of the visit
Sunshine 7 days	The average daily minutes of sunshine in the last 7 days before the visit occasion
Sunshine 30 days	The average daily minutes of sunshine in the last 30 days before the visit occasion
Temperature	The mean temperature on the day of the visit
Temperature 7 days	The average temperature in the last 7 days before the visit occasion
Temperature 30 days	The average temperature in the last 30 days before the visit occasion
Time variables:	
Time morning	A binary variable indicating whether the customer will be visited in the morning (before 1 p.m.)
Time afternoon	A binary variable indicating whether the customer will be visited in the Afternoon (between 1 p.m. and 5 p.m.)
Time evening	A binary variable indicating whether the customer will be visited in the evening (after 5 p.m.)
Sales person variables:	
Salesperson	A categorical variable indicating the sales person

Table 1: Model variables

Physical surroundings are represented by weather variables, more specifically by the minutes of sunshine and the mean temperature. Besides the weather condition during the purchase occasion, also historical weather information of the last seven and thirty days before the purchase occasion will be incorporated. As temporal perspective the moment of the day is included. Because a salesperson cannot always follow the schedule very strictly, the actual visit time can sometimes differ from the scheduled one. Hence, it is preferable to create a time variable that is not too detailed, such as the moment of the day, consisting of morning, afternoon and evening. The most important social surrounding is the influence of the salesperson who visits the client. Because every one of the 175 salespeople in this model has unique attitudinal and behavioral characteristics, correlation between the outcomes of the purchase occasions of the same salesperson can be expected. Therefore, a multilevel model based on this variable is introduced to capture this effect. This research will first investigate data augmentation with each of the three situational variables added one by one. Next, a final model will be composed including all transactional and situational predictors.

3.2. Classification techniques

Modeling whether a visited customer will purchase at least one product, results in a binary classification problem. This paragraph introduces two statistical techniques used throughout this study that are able to handle such problems. The basic model and the models augmented with weather and time variables are based on logistic regression techniques. In order to capture the salesperson effect a multilevel model is introduced.

3.2.1. Logistic regression model

Logistic regression is a well-known technique frequently used in traditional marketing applications.[40] An important benefit over other methods (e.g. neural networks) is its interpretability. It produces specific information about the size and direction of the effects of independent variables. Moreover, in terms of predictive performance and robustness, logistic regression can compete with more advanced data mining techniques.[41] Logistic regression belongs to the group of generalized linear models (GLM). GLMs adopt ordinary least square regression to other response variables, like dichotomous outcomes, by using a link function[42]. In logistic regression the parameters are estimated by maximizing the log-likelihood function. Including these estimates in the following formulae creates probabilities, ranging from 0 to 1, that can be used to rank customers in terms of their likelihood of purchase.[43]

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{1}$$

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} \tag{2}$$

Whereby: π_i represents the *a posteriori* probability of purchase by customer i; X_{ni} represents the independent variables for customer i; β_0 represents the intercept; β_n represent the parameters to be estimated; n represents the number of independent variables.

Due to the high correlation between independent variables, it is possible that some variables, although significant in a univariate relationship, have little extra predictive value to add to the model. Hence, this study will include a backward selection technique that creates a subset of the original variables by eliminating variables that are either redundant or possess little additional predictive information. This should enhance the comprehensibility of the model and decrease the computation time and cost, which is very important in a highly dynamic model that must be scored on a daily basis.[25]

3.2.2. Multilevel model

Originally, multilevel or hierarchical models were often used in research disciplines as sociology to analyze a population structured hierarchically in groups or clusters. For example, in Ref 44 students on the lowest level are nested within schools on a higher level. In such samples, the individual observations are often not completely independent. As a result, the average correlation between variables measured on observations within the same group will be higher than the average correlation between variables measured based on observations from different groups. Standard statistical techniques, such as logistic regression, rely heavily on the assumption of independence of observations and a violation of this assumption can have a significant influence on the accuracy of the model.[45] In this study it is expected that due to the differences in personal attitudinal and behavioral characteristics between salespersons, purchase occasions of the same salesperson will have a higher correlation than average. In other words, purchase occasions can be nested within salespeople.

There are several ways to extend a single-level model to a multilevel model. The easiest way to take the effects of higher-level units into account is by adding dummy variables so that each higher-level unit has its own intercept in the model. These dummy variables can be used to measure the differences between salespersons. The use of fixed intercepts, however, increases the number of additional parameters equal to the number of higher-level units minus one. Because this study includes 175 salespeople, this would result in a large number of nuisance parameters in the model. A more sophisticated approach is to treat the salesperson intercepts as a random variable with a specified probability distribution in a multilevel model. This method will lead to more accurate predictions.

Assuming that data is available from J groups with a different number of observations n_j in each group, a multilevel model can be estimated based on the following equation:[45]

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \tag{3}$$

In this equation, Y_{ij} and X_{ij} represent the dependent and one (or more) independent variables at the lowest level respectively. The residual errors e_{ij} are assumed to be normally distributed with a mean of zero and a variance, denoted by σ_e^2, that has to be estimated. The intercept and slope coefficients, β_{0j} and β_{1j} respectively, are assumed to vary across the groups. These coefficients, often called random coefficients, have a distribution with a certain mean and variance that can be explained by one or more independent variables at the highest level Z_j, as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \tag{4}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} \tag{5}$$

The u-terms u_{0j} and u_{1j} represent the random residual errors at the highest level and are assumed to be independent from the residual errors e_{ij} at the lowest level and normally distributed with a mean of zero and a variance of $\sigma_{u_0}^2$ and $\sigma_{u_1}^2$ respectively. The covariance between the residual error terms u_{0j} and u_{1j}, denoted as $\sigma_{u_{01}}^2$, is generally not assumed to be zero.

By substituting “Eq. (4)” and “Eq. (5)” into equation “Eq. (3)” and rearranging terms, a single complex multilevel equation is created:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\, X_{ij} + \gamma_{01}\, Z_j + \gamma_{11}\, Z_j\, X_{ij} + u_{1j}\, X_{ij} + u_{0j} + e_{ij} \tag{6}$$

This model can be split into a fixed or deterministic part $[\gamma_{00} + \gamma_{10}\, X_{ij} + \gamma_{01}\, Z_j + \gamma_{11}\, Z_j\, X_{ij}]$ and a random or stochastic part $[u_{0j} + u_{1j}\, X_{ij} + e_{ij}]$. This illustrates that, in order to allow correlation between the observations, the generalized linear model (GLM) must be extended to a generalized linear mixed model (GLMM) with random effects that are assumed to be normally distributed.

In our study the dependent variable at the lowest level is the outcome whether the client purchased at least one product during the purchase occasion. Because this is a dichotomous variable, "Eq. (6)" needs to be transformed using a logit link function in the following way:[45]

$$Y_{ij} = \pi_{ij}\,;\ \pi \sim \text{Binomial}(n_{ij}, \mu) \tag{7}$$

$$\pi_{ij} = \text{logistic}(\,\gamma_{00} + \gamma_{10}\, X_{ij} + \gamma_{01}\, Z_j + \gamma_{11}\, Z_j\, X_{ij} + u_{1j}\, X_{ij} + u_{0j}) \tag{8}$$

These equations state that the dependent variable is a proportion π_{ij} , assuming to have a binomial error distribution with sample size n_{ij} and expected value μ. If all possible outcomes are only zero and one, the sample sizes are reduced to one and dichotomous data is modeled. Due to the binomial distribution, the lowest-level residual variance is a function of the proportion:

$$\sigma_e^2 = \frac{\pi_{ij}}{1-\pi_{ij}} \tag{9}$$

	Logistic regression model	**Logistic multilevel model**
Model family:	Generalized linear model (GLM)	Generalized linear mixed model (GLMM)
Regression equation:	$Y_i = \beta_0 + \beta_1 X_i + e_i$	$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$
Link function for dichotomous outcomes:	$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$	$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$
Correlation between observations:	Not assumed	Allowed
Relationship between dependent and independent variables:	Assumed to be linear	Assumed to be linear

Table 2: Comparison between a logistic regression model and a logistic multilevel model

Consequently, this variance does not have to be estimated separately and the lowest-level residual errors e_{ij} can be excluded from the equation. In Table 2 a summarized comparison between a logistic regression model and a logistic multilevel model can be found.

The database from this study does not contain meaningful higher-level information about the salespeople. Furthermore, it is not expected that the slopes of any of the lower-level variables will vary across the salespeople. This makes it possible to reduce “Eq. (8)” to:

$$\pi_{ij} = \text{logistic}(\, \beta_{0j} + \beta_1 X_{ij}) \tag{10}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{11}$$

Combining “Eq. (10)” and “Eq. (11)” results into:

$$\pi_{ij} = \text{logistic}(\,\gamma_{00} + \beta_1 X_{ij} + u_{0j}\,) \tag{12}$$

This hierarchical logistic regression model still contains a fixed part [$\gamma_{00} + \beta_1 X_{ij}$] and a random part [$u_{0j}$].

The intraclass correlation coefficient (ICC), which measures the proportion of variance in the outcome explained by the grouping structure, can be calculated using an intercept-only model. This model can be derived from “Eq. (8)” by excluding all explanatory variables, which results in the following equation:

$$\pi_{ij} = \text{logistic}(\,\gamma_{00} + u_{0j}) \tag{13}$$

The ICC is then calculated based on the following formula:[45]

$$\text{ICC} = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2} \tag{14}$$

Because the variance of a logistic distribution with scale factor 1 is π²/3 ≈ 3.29 in a hierarchical logistic regression model, this formula can be reformulated as:[45]

$$\text{ICC} = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + {}^{\pi^2}\!/_{3}} \tag{15}$$

3.3. Evaluation criterion

In order to be able to evaluate the predictive performance of each model the database, containing 162,424 observations, is randomly split into two equal parts. The first part, called training sample, is used to estimate the model. Afterwards, this model is validated on the remaining 50% of observations. It is essential to evaluate the performance of the classifiers on a holdout validation sample in order to ensure that the training model can be generalized over all customers of the home vending company. The analysis table is generated based on transactional information during the period between February 1st, 2007 and November 30th, 2007. Besides the training and validation sample, also an out-of-period test sample is created based on the same period in 2008, containing 161,462 observations. Using the model trained on data of 2007, predictions are made for all observations in the out-of-period test sample. This makes it possible to check the evolution of the accuracy of the model over time. If the performance does not drop significantly, the model can be generalized not only over all customers of the home vending company, but also over different time periods.

The area under the receiver operating characteristic curve (AUC) is used as evaluation metric of the classifiers.[46] The advantage of an AUC in comparison with other evaluation metrics, like the percent correctly classified (PCC), is the fact that PCC is highly dependent on the chosen threshold that has to be determined to distinguish the predicted events from non-events. The calculation of the PCC is based on a ranking of customers according to their *a posteriori* probability of purchase. Depending on the context of the problem of the home vending company (e.g. the amount of the capacity problem) a cutoff value is chosen. All customers with an *a posteriori* probability of purchase higher than the cutoff are classified as buyers and will be visited. All customers with a lower likelihood of purchase are labeled as non-buyers. This classification can be summarized in a confusion matrix, displayed in Table 3.[47]

		Predicted status	
		Buyer	**Non-buyer**
True Value	**Buyer**	True Positive (TP)	False Negative (FN)
	Non-buyer	False Positive (FP)	True Negative (TN)

Table 3: Confusion matrix

Based on this matrix the percentage of correctly classified observations can be formulated as:[48]

$$\text{PCC} = \frac{\text{TP+TN}}{\text{TP+ TN+ FP+FN}} \tag{16}$$

Besides the PCC, the following meaningful measures can also be calculated:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}} \tag{17}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{18}$$

Sensitivity represents the proportion of actual events that the model correctly predicts as events (i.e. the number of true positives divided by the total number of events). Specificity is defined as the proportion of non-events that are correctly identified (i.e. the number of true negatives divided by the total number of non-events). It is important to notice that all these measures give only an indication of the performance at the chosen cutoff. In reality, the chosen cutoff will vary depending on the context of the problem of the decision maker, hence an evaluation criterion independent of the chosen cutoff, such as the AUC, is preferred.

The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of sensitivity and one minus specificity for all possible cutoff values used (e.g. Fig. 1). The AUC measures the area under this curve and can be interpreted as the probability that a randomly chosen positive instance is correctly ranked higher than a randomly selected negative instance.[46] This again illustrates that this evaluation criterion is independent of the chosen threshold. As a result, this criterion is often used as evaluation metric for the predictive performance of CRM models (e.g. Ref. 29). The AUC measure can range from a lower limit of 0.5, if the predictions are random (corresponding with the diagonal in Fig. 1), to an upper limit of 1, if the model's predictions are perfect.

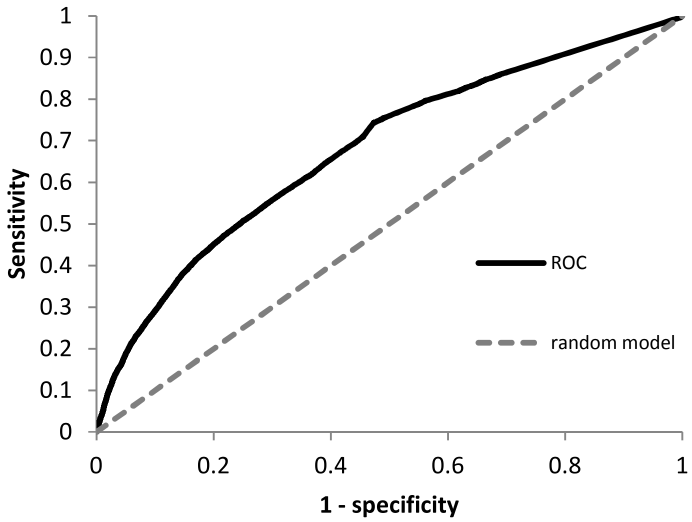


Figure 1: AUC example

4. Results

The results of this study are clearly summarized in Table 4 and Table 5. In Table 4 all parameter estimates of each model are described. First, the basic model, based on only transactional data, will be discussed. Next, this model will be enhanced with each of the situational variables added one by one in order to examine the individual effect. Eventually, all variables will be incorporated in a final model. In this table, only the significant variables after the backward selection technique are retained. Because of the high number of observations, a significance level of 0.01 is preferred. In Table 5 the predictive performance, in terms of AUC, is displayed for the training, validation and out-of-period test sample.

	Logistic regression model						**Multilevel model**			
	Basic model		**+ Weather variables**		**+ Time variables**		**+ Salesperson variables**		**Final model**	
Variable	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	-0.7425	0.0344	-1.0792	0.0386	-0.7604	0.0345	-0.6460	0.0447	-0.9875	0.0480
Transactional variables:										
Recency visit	0.0062	0.0008	0.0060	0.0008	0.0060	0.0008	0.0023	0.0008	0.0021	0.0008
Frequency bought	0.4031	0.0184	0.3552	0.0186	0.4055	0.0184	0.4959	0.0194	0.4448	0.0197
Sales ratio	1.0153	0.0650	1.1570	0.0658	1.0091	0.0650	0.5762	0.0693	0.7339	0.0700
Avg. mon. value	0.0115	0.0021	0.0115	0.0021	0.0113	0.0021	0.0139	0.0021	0.0136	0.0021
Last time visit	-0.2697	0.0246	-0.2610	0.0248	-0.2683	0.0246	-0.2639	0.0255	-0.2510	0.0257
Last time bought	-0.5984	0.0221	-0.6020	0.0222	-0.5994	0.0221	-0.6158	0.0223	-0.6195	0.0224
Weather variables:										
Sunshine			0.0002	0.0000					0.0002	0.0000
Sunshine 7 days			0.0007	0.0001					0.0007	0.0001
Sunshine 30 days			0.0004	0.0001					0.0003	0.0001
Time variables:										
Time evening					0.1346	0.0203			0.0650	0.0216
Salesperson variables:										
Intercept variance ($\sigma_{u_0}^2$)							0.1208	0.0151	0.1171	0.0146

Table 4: Overview of the parameter estimates

Sample	**Basic model**	**+ Weather variables**	**+ Time variables**	**+ Salesperson variables**	**Final model**
Training sample	0.6793	0.6861	0.6804	0.7014	0.7054
Validation sample	0.6801	0.6871	0.6816	0.6996	0.7039
Out-of-period test sample	0.6818	0.6885	0.6837	0.6996	0.7035

Table 5: Model performance measured in term of AUC

4.1. Basic model

A logistic regression model that only uses transactional variables in order to predict purchasing behavior will be used as benchmark model. Because of the backward selection technique, only six of the initial ten input variables are retained. High correlation between some of the transactional variables results in the fact that four variables do not add extra predictive value to the model. Having a closer look at the parameter estimates in Table 4 gives interesting insights into the purchasing pattern of the home vending company's customers. All significant variables based on the past purchasing behavior in the last eight weeks (i.e. frequency bought, sales ratio and average monetary value) have a positive relationship with the future purchasing behavior. On the other hand, the transactional variables based on the last visit (i.e. last time visit and last time bought) all have a negative relationship with the probability to purchase on a next visit. Normally, a customer is visited in a biweekly schedule. This means that, if there are no capacity problems, there are 14 days between visiting the same customer again. These parameter estimates imply that the most attractive customers have high RFM scores in general, but if the customer was visited at a normal frequency the last time and moreover bought a product, his/her probability of buying the next time will drop. Although, if a customer was not visited (e.g. due to capacity problems) the dummy variables last time visit and last time bought will be flagged zero, as a result his/her probability to purchase next time will rise and the chance that (s)he will be excluded again will decrease. This illustrates the usefulness of a dynamic model that ranks customers on a daily basis in order to ensure that, at every moment, priority is given to clients with the highest purchase probability. With an AUC of 0.6793, 0.6801 and 0.6818 on the training, validation and out-of-period test sample respectively (Table 5), this study confirms that variables about the past purchasing behavior are still good predictors for future purchasing behavior. Notwithstanding this relative good performance based on transactional data, improvement can still be obtained by data augmentation with situational variables.

4.2. Data augmentation with weather variables

Besides transactional data, enhancing a database with physical surroundings in the form of weather variables can improve the accuracy of a purchase prediction model. This study incorporates sunshine and temperature, but Table 4 illustrates that only the sunshine variables are significantly related to purchasing. Actually, in a univariate relationship, temperature is also significant, but it does not deliver extra predictive value on top of the other variables. Table 5 indicates that on the three samples used in this study, a significant improvement in terms of AUC is found by taking sunshine variables into account.

4.3. Data augmentation with time variables

The temporal perspective is a second situational dimension that can be used for data augmentation. This study investigates the effect of including the moment of the day that the salesperson will visit the customer on the predictive performance of the model. Table 4 indicates that visiting customers after 5 p.m. increases the probability of purchase. An explanation for this phenomenon cannot be found in the fact that most people are at work before 5 p.m. because this model captures only observations where the client was at home. One possible explanation can be found in the literature of time pressure. Ref. 38 already demonstrates that time pressure has a negative effect on purchasing behavior. The assumption that people experience less time pressure at the end of the day can be an explanation for the positive relationship between evening visits and purchasing behavior. No significant differences were found for visits at the morning or afternoon. Adding this single dummy variable to the basic model, results in a small, but still significant increase in predictive performance (Table 5).

4.4. Data augmentation with salesperson variables

In order to take the effect of social surroundings into account, a multilevel model is introduced. In this study the most important social surrounding at the purchase occasion is the personal influence of one of the 175 salespeople. First, the intraclass correlation coefficient is calculated based on an intercept-only model without independent variables. In this model, the intercept variance ($\sigma_{u_0}^2$) was estimated to be 0.1716. Using formula (15), this results in an ICC of 0.0496, meaning that 4.96% of the variation in the purchasing behavior can be explained by grouping the customers based on the salespeople who visit them. Table 5 indicates that by structuring the purchase occasions by salesperson a strong increase in predictive performance can be obtained using the same transactional variables, can be obtained. Furthermore, it should be noticed that the estimate of the intercept variance drops to 0.1208 due to the inclusion of independent transactional variables in the model (Table 4).

4.5. Final model

Data augmentation with each of the three groups of situational variables resulted in a higher predictive performance on the training, validation and out-of-period test sample. All pairwise comparisons of all models reported in Table 5 resulted in significant differences based on the non-parametric test of Delong *et al.*[49] The most improvement in predictive performance was obtained by taking the salesperson effect into account. The second largest increase in AUC results from the enhancement of the database with three sunshine variables. Furthermore, taking into account that evening visits are positively related with purchase also leads to a small, but still significant improvement in accuracy. Eventually, all variables are incorporated in a final model. Table 4 indicates that in this model all relationships remain significant at a 0.01 significance level. This implies that the three groups of situational variables each explain a different part of the variance in purchasing behavior. A comparison between the predictive performance of the final model and the basic model in Table 5 shows that data augmentation with situational variables can be very useful to identify the customers with the highest probability of purchase. This study is able to improve the AUC by 0.0261, 0.0238 and 0.217 on the training, validation and out-of-period test sample respectively. Differences between the AUCs of the three samples are relatively small, which implies that this model can be generalized over time and to all customers of the home vending company.

5. Conclusion and Further Research

In order to remain competitive, a lot of firms implement information technology tools to improve their marketing strategies.[50, 51] Nowadays, an increasing number of software products are available to support decision making.[52] As a result, the company's database has become a valuable asset to support marketing decisions. Also academic researchers constantly try to improve CRM models in general, and predictive analytics in particular. This is possible by focusing on the data mining techniques, but the enhancement of the database itself, on which these data mining techniques are run, can also result in improved predictive performance of CRM models. This study suggests not to restrict the predictors of a CRM model to variables that are only related to the individual (e.g. the individual past purchasing behavior). Taking into account the situational information about the purchase occasion can significantly improve purchasing behavior predictions.

For a home vending company, some of the situational information is known in advance and can easily be included in a highly dynamic model that scores the customers on a daily basis. Three dimensions of situational variables were examined: physical surroundings, temporal perspective and social surroundings. A small, but still significant improvement in accuracy was observed by data augmentation with the temporal perspective dimension. Higher probabilities to purchase are estimated when a salesperson visits the customer in the evening. Probably, customers experience less time pressure at the end of the day and consequently are more willing to purchase. Based on these findings, the home vending company can try to shift the working hours of his salespeople more towards the evening in order to improve the sales ratio.

Besides the moment of the day, the incorporation of physical surroundings in the form of weather variables was inspected. Although temperature was significant in a univariate relationship, only the sunshine variables were able to add extra predictive value to the model. By combining these findings with weather forecast predictions, the home vending company should be able to better foresee and manage capacity problems. This model, in combination with flexible salespeople, can be used by a marketing decision maker to shift more resources to periods with higher purchase probabilities. If the demand is still too high to visit every customer, the model can be used to give priority to customers with the highest probability to buy.

The best increase in predictive performance was obtained by taking social surroundings, represented by the salesperson effect, into account using a multilevel model. Fig. 2. represents the intercepts for each of the 175 salespeople estimated by the final multilevel model. The values are ranked from lowest on the left side to highest on the right side. This figure illustrates that attitudinal and behavioral differences between salespeople result in a significant variation in the ability to sell products. Hence, the home vending company should take these intercept estimations into account during the evaluation process of its salespeople.

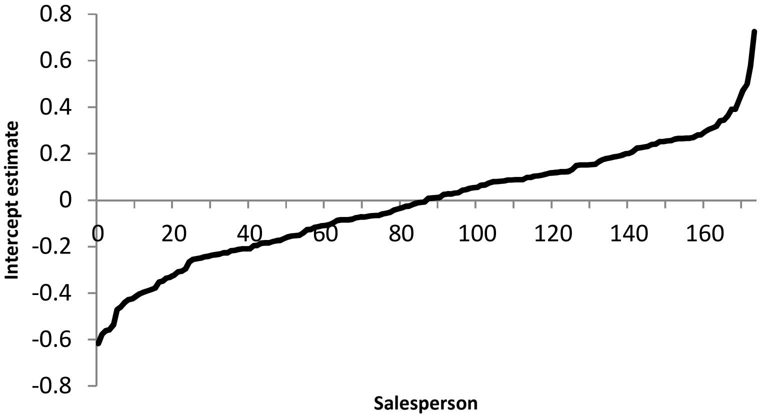


Figure 2: Intercept estimates for each salesperson

In a final model, all variables are included resulting in a significant, but also economically relevant improvement of predictive performance.

While this study fills a gap in today's literature by using situational variables for data augmentation in a CRM context, there are still some recommendations for further research. It should be mentioned that this analysis is done in a specific setting based on data of a home vending company, specialized in frozen foods and ice cream, to predict purchasing behavior. In order to be able to generalize the findings of this study, similar analyses in a different framework, should be conducted. Furthermore, the situational variables are not restricted to the ones described in this research. Further research could investigate if there are still other undiscovered situational variables that can be considered for data augmentation. In this study, we found evidence that customers are more willing to purchase in the evening. A probable explanation could be that people feel less time pressure at the end of the day and as a result are more willing to purchase. Only the relationship between time pressure and purchasing behavior has already been investigated,[38] but, to the best of our knowledge, no research is found about the relationship between time pressure and the moment of the day.

Acknowledgement

Both authors acknowledge the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

References

S. Lipovetsky, SURF - Structural Unduplicated Reach and Frequency: Latent class TURF and Shapley Value analyses, *Int. J. Inf. Technol. Decis. Mak.* **7**(2008) 203-216.

R. Ling, and D. C. Yen, Customer relationship management: An analysis framework and implementation strategies, *Journal of Computer Information Systems* **41**(2001) 82–97.

D. Van den Poel, and W. Buckinx, Predicting online-purchasing behavior, *Eur. J. Oper. Res.* **166**(2005) 557–575.

R. Al-Aomar, and F. Dweiri, A customer-oriented Decision Agent for product selection in web-based services, *Int. J. Inf. Technol. Decis. Mak.* **7**(2009) 35-52.

L. A. Petrison, R. C. Blattberg, and P. Wang, Database marketing past, present and future, *Journal of direct marketing* **7**(1993) 27-43.

E. W. T. Ngai, L. Xiu, and D. C. K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Syst. Appl.* **36**(2009) 2592-2602.

W. Kamakura, C. F. Mela, A. Ansari, A. Bodapati, P. Fader, R. Iyengar, P. Naik, S. Neslin, B. Sun, P.C. Verhoef, M. Wedel, and R. Wilcox, Choice Models and Customer Relationship Management, *Mark. Lett.* **16**(2005) 279-291.

R. Khan, M. Lewis, and V. Singh, Dynamic Customer Management and the Value of One-to-One Marketing, *Mark. Sci.* **28**(2009) 1063-1079.

A. Krasnikov, S. Jayachandran, and V. Kumar, The Impact of Customer Relationship Management Implementation on Cost and Profit Efficiencies: Evidence from the US Commercial Banking Industry, *J. Mark.* **73**(2009) 61-76.

T. S. H. Teo, P. Devadoss, and S. L. Pan, Towards a holistic perspective of customer relationship management implementation: A case study of the housing and development board, Singapore, *Decis. Support Syst.* **42**(2006) 1613–1627.

P. Baecke, and D. Van den Poel, Data augmentation by predicting spending pleasure using commercially available external data, *J. Intell. Inf. Syst.* **36**(2010) 367-383.

W. Buckinx, E. Moons, D. Van den Poel, and G. Wets, Customer-adapted coupon targeting using feature selection, *Expert Syst. Appl*. **26**(2004) 509–518.

K. A. Smith, R. J. Wills, and M. Brooks, An analysis of customer retention and insurance claim patterns using data mining: A case study, *J. Oper. Res. Soc.* **51**(2000) 532–541.

S. Gupta, D. R. Lehmann, and J. A. Stuart, Valuing customers, *J. Mark.* **41**(2004) 7–19.

F. F. Reichheld, and W. E. Sasser, Zero defections: Quality comes to services, *Harv. Bus. Rev.* **68**(1990) 105–112.

A. Prinzie, and D. Van den Poel, Random Forests for Multiclass classification: Random Multinomial Logit, *Expert Syst. Appl.* **34**(2008) 1721-1732.

B. Boutsinas, and S. Athanasiadis, On merging classification rules, *Int. J. Inf. Technol. Decis. Mak.* **7**(2008) 431-450.

Y. Peng, G. Kou, Y. Shi, and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, *Int. J. Inf. Technol. Decis. Mak.* **7**(2008) 639-682.

C. Hung, and C. Tsai, Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand, *Expert Syst. Appl.* **34**(2008) 780–787.

E. H. Suh, K. C. Noh, and C. K. Suh, Customer list segmentation using the combined response model, *Expert Syst. Appl.* ***17***(1999) 89–97.

J. R. Bult, and T. Wansbeek, Optimal selection for direct mail, *Mark. Sci.* **14**(1995) 378–394.

J. A. McCarty, and M. Hastak, Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, *J. Bus. Res.* ***60***(2007) 656–662.

H. Shin, and S. Cho, Response modeling with support vector machines, *Expert Syst. Appl.* ***30***(2006) 746-760.

J. Zahavi, and N. Levin, Applying neural computing to target marketing, *Journal of Direct Marketing* **11**(1997) 5–23.

Y. S. Kim, Toward a successful CRM: Variable selection, sampling, and ensemble, *Decis. Support Syst.* **41**(2006) 542–553.

C. H. Cheng, and Y. S. Chen, Classifying the segmentation of customer value via RFM model and RS theory, *Expert Syst. Appl.* **36**(2009) 4176–4184.

T. J. Steenburgh, A. Ainsle, and P. H. Engbretson, Massively Categorical Variables, Revealing the Information in ZIP-Codes, *Mark. Sci.* **22**(2003) 40–57.

J Hu, and N. Zhong, Web farming with clickstream, *Int. J. Inf. Technol. Decis. Mak.* **7**(2008) 291-308.

S. Hill, F. Provost, and C. Volinsky, Network-based marketing: Identifying likely adopters via consumer networks, *Stat. Sci.* **21**(2006) 256-276.

C. Coussement, and D. Van den Poel, Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers, *Expert Syst. Appl.* **36**(2009) 6127–6134.

T. S. Lix, P.D. Berger, and T.L. Magliozzi, New customer acquisition: Prospecting models and the use of commercially available external data, *Journal of Direct Marketing* ***9***(1995) 8–19.

R. W. Belk, Situational Variables and Consumer Behavior, *J. Consum. Res.* **2**(1975), 157-164.

S. Roslow, T. Li, and J. Nicholls, Impact of situational variables and demographic attributes in two seasons on purchase behaviour, *European J. Mark.* **34**(2000) 1167-1180.

J. J. Denissen, L. Butalid, L. Penke, and M. A.Van Aken, The Effects of Weather on Daily Mood: A Multilevel Approach, *Emotion* **8**(2008) 662-667.

D. Hirshleifer, and T. Shumway, Good Day Sunshine: Stock Returns and the Weather, *Journal of Finance* **58**(2001) 1009-1032.

O. Levy, and I. Galili, Stock purchase and weather: Individual differences, *J. Econ. Behav. Organ.* **67**(2008) 755-767.

E. M. Saunders, Stock Prices and Wall Street Weather, *Am. Econ. Rev.* **83**(1993) 1337-1345.

C. W. Park, E. S. Iyer, and D. C. Smith, The Effect of Situational Factors on In-store Grocery Shopping Behavior: The Role of Store Environment and Time Available for Shopping, *J. Consum. Res.* **15**(1989) 422-433.

G. Albaum, Exploring Interaction in a Marketing Situation, *J. Mark. Res.* **4**(1967) 168-72.

R. E. Bucklin, and S. Gupta, Brand choice, purchase incidence and segmentation: An integrated modeling approach, *J. Mark. Res.* **29**(1992) 201–215.

N. Levin, and J. Zahavi, Continuous predictive modeling: A comparative analysis, *J. Interact. Mark.* **12**(1998) 5–22.

P. McCullagh and J. A. Nelder, *Generalized linear models (second edition)* (Chapman & Hall, London, 1989).

D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression (second edition) (*John Wiley & Sons, New York, 2000).

V. E. Lee, and A. S. Bryk, A multilevel model of the social distribution of high school achievements, *Sociol. Educ.* **62**(1989) 172-192.

J. Hox, *Multilevel Analysis: Techniques and Applications* (Taylor & Francis Group, New York, 2002).

J. A. Hanley, and B. J. McNeil, The meaning and use of area under a receiver operating characteristic (ROC) curve, *Radiology* **143**(1982) 29–36.

CHAPTER IV:

IMPROVING CUSTOMER ACQUISITION MODELS BY INCORPORATING SPATIAL AUTOCORRELATION AT DIFFERENT LEVELS OF GRANULARITY

This chapter is based upon Baecke, P. & Van den Poel, D. (2012). Improving Customer Acquisition models by Incorporating Spatial Autocorrelation at Different Levels of Granularity, *under review at the Journal of Intelligent Information Systems.*

CHAPTER IV:
IMPROVING CUSTOMER ACQUISITION MODELS BY INCORPORATING SPATIAL AUTOCORRELATION AT DIFFERENT LEVELS OF GRANULARITY

Abstract

Traditional CRM models often ignore the correlation that could exist among the purchasing behavior of surrounding prospects. Hence, a generalized linear autologistic regression model can be used to capture this interdependence and improve the predictive performance of the model. In particular, customer acquisition models can benefit from this. These models often suffer from a lack of data quality due to the limited amount of information available about potential new customers. Based on a customer acquisition model of a Japanese automobile brand, this study shows that the extra value resulting from incorporating neighborhood effects can vary significantly depending on the granularity level on which the neighborhoods are composed. A model based on a granularity level that is too coarse or too fine will incorporate too much or too little interdependence resulting in a less than optimal predictive improvement. Since neighborhood effects can have several sources (i.e. social influence, homophily and exogeneous shocks), this study suggests that the autocorrelation can be divided into several parts, each optimally measured at a different level of granularity. Therefore, a model is introduced that simultaneously incorporates multiple levels of granularity resulting in even more accurate predictions. Further, the effect of the sample size is examined. This showed that including spatial interdependence using finer levels of granularity is only useful when enough data is available to construct stable spatial lag effects. As a result, extending a spatial model with multiple granularity levels becomes increasingly valuable when the data sample becomes larger.

1. Introduction

Customer Relationship Management (CRM) has become an important topic in the field of marketing [1]. The technological development, the rise of the internet and declining costs for data warehousing and information processing have encouraged companies to collect data about their customers and prospects [2]. CRM uses data mining techniques to convert this unstructured data into valuable information. This has resulted in the development of useful information technology tools to support marketing decision making and predict the effect of it [3,4].

Besides the data mining technique, the success of a CRM model also depends on the quality of the information used as input for the model [5]. Traditional CRM models often ignore neighborhood information and rely on the assumption of independent observations. This means that customers' purchasing behavior is totally unrelated to the behavior of others. However, in reality, customer preferences do not only depend on their own characteristics, but are often also related to the behavior of other customers in their neighborhood. Using neighborhood information to incorporate spatial autocorrelation in the model can solve this shortcoming and significantly improve the predictive performance of the model.

Several studies have already proven that spatial statistics can produce interesting insights in marketing [6-12]. However, only a limited number of studies use spatial information to improve the accuracy of a predictive CRM model. In reference [13], customer interdependence was estimated based on geographic and demographic proximity. The study indicated that geographic reference groups are more important than demographic reference groups in determining individual automobile preferences. Reference [14] showed that taking zip-code information into account can significantly improve a model used for the attraction of new students by a private university. The focus of this research will also be only on physical geographic interdependence, but compared to previous literature, this study includes a high number of independent socio-demographic and lifestyle variables that are typically available at an external data vendor. This should prevent the predictive improvement to be caused by the absence of other important variables that can be easily obtained for customer acquisition models.

In this paper, neighborhood information is used to incorporate spatial autocorrelation in a customer acquisition model for a Japanese car brand. Reference [15] is the first paper that compared the value of incorporating spatial information in CRM models across multiple product categories. That study found that especially for publicly consumed durable goods, such as automobile brands, incorporating neighborhood effects can be very useful. Further, within CRM models, customer acquisition models suffer the most from a lack of data quality. A company’s customer database is typically single source in nature. The data collection is limited to the information a company retrieves from its own customers. As a result, for customer acquisition campaigns the company has to attract data from external data vendors. Nevertheless, these data still only contains a limited number of socio-demographic and lifestyle variables [16]. Especially in such situation, incorporating extra neighborhood information can improve the identification of potential customers.

In addition, extra complexity is introduced that has been mostly ignored in previous literature. Customers can often be clustered in neighborhoods at multiple levels (e.g. country, district, ward, etc.). In order to incorporate these neighborhood effects efficiently, the level of granularity should be carefully chosen. If the neighborhood is chosen too large, the spatial interdependence will fade away because the preferences of too many surrounding customers are taken into account that do not have any influence in reality. On the other hand, choosing neighborhoods that are too small can affect the stability of the measured influence and ignore the correlation with some customers that still have an influence. This study will compare the relevance of taking neighborhood effects into account at different levels of granularity.

In order to facilitate the decision making about the optimal granularity level, a model is introduced that simultaneously incorporates multiple levels. Such a model is developed based on the assumption that multiple sources are responsible for the existence of autocorrelation between customers' purchasing behaviors and each of these sources will have a different range in which interdependence exists. As a result, this model is able to incorporate spatial autocorrelation from several sources, each at their optimal granularity level.

Furthermore, this study will investigate how the size of the dataset can influence the predictive performance of the spatial models. These spatial models take the purchasing behavior of surrounding customers into account to assist in purchasing behavior predictions of a particular customer. At a finer level of granularity, customers are divided into more neighborhoods in which spatial interdependence is assumed. As a result, only closer neighbors, who are assumed to have a higher influence, are used to assist in the predictions. On the other hand though, this also results in fewer observations available to construct these spatial influences, which may affect the stability of the spatial variables. Consequently, increasing the data sample should improve the incorporation of spatial interdependence calculated on finer granularity levels.

The remainder of this paper is organized as follows. Section 2 will elaborate on several sources that are responsible for the existence of spatial interdependence in CRM models. The methodology is described in Section 3, consisting of the data description, the generalized linear autologistic regression model and the evaluation criterion used in this study. The results are reported in Section 4 and Section 5 provides a discussion of these results in combination with a conclusion.

2. Origins of spatial interdependence

In this study neighborhood effects are defined as the existence of correlating purchasing behavior among geographically closely located customers. Based on previous literature, three concepts can be distinguished that are responsible for the existence of this spatial interdependence, namely social influence, homophily and exogenous shocks. The focus of this study is not to disentangle the effect of these three concepts, but to simultaneously take all these effects into account in order to obtain more accurate CRM models.

In the following sections these concepts are described, illustrating that the spatial autocorrelation caused by each effect may be optimally measured at different granularity levels. Hence, the added value of incorporating interdependence in a customer acquisition model can differ significantly depending on the granularity level that is used to compose the neighborhoods. Furthermore, a generalized linear autologistic regression model that allows dividing the spatial autocorrelation over multiple granularity levels can improve predictions even more.

2.1. Social influence

The power of social influence in marketing has been known for some time [17]. Customers do not live in an isolated environment where decisions are made in a purely rational way. Instead, product preferences and purchasing decisions are often influenced by positive and negative recommendations of other individuals. Word of mouth (WOM) can have an important impact on a customer's decision because this information is perceived as highly credible [18]. Due to its non-commercial nature this information is processed with less skepticism than advertising or promotion. Although the emergence of online word of mouth should not be ignored, the majority of word-of-mouth conversations still take place in face-to-face interpersonal settings. More specifically, Reference [19] and [20] show that still 76% to 80% of the WOM conversations occur face-to-face, while only about 10% are online. Further, it can be assumed that people who live in the same neighborhood will have more correlated purchasing behavior, as living closer together provides more opportunities for interaction and communication. This has also been supported by reference [12] in which spatial proximity is used as proxy for WOM to investigate contagion in new product adoption. As a result, geographic proximity can still be considered as an important indication of social influence. Although online product recommendations will also have an influence on the purchasing behavior of the customers, already a large part of this social influence can be taken into account by using geographical information. In addition, spatial variables are ideal for data augmentation applications since these can be easily collected for a large number of customers.

Actually, customers do not even have to interact to affect each other. Observing the purchasing decisions of others can be enough to influence an individual's purchasing decision [21]. In other words, besides WOM, observational learning (OL) is a second important social influence that can be responsible for spatial autocorrelation in a CRM model. Neighboring customers buy similar products and brands not only because they want to match the social standard of the neighborhood, but also because they may be more confident about the quality if they recognize that many people bought the product or brand. Although WOM contains more information because it makes it possible to clarify an opinion or recommendation, the information from OL might be perceived as more credible because it reveals the real action of other consumers [22].

2.2. Homophily

Besides social influences, another explanation of the existence of interdependence between customers' purchasing behavior is homophily, also called endogenous group formation [23]. This concept is often referred to with the proverbial expression "Birds of a feather flock together" [24].

In other words, people with similar tastes and characteristics tend to group together. Two types of homophily can be distinguished to explain the existence of sociospatial patterns, namely social homophily and structural homophily [25]. Social homophily means that people wish to live close to others with similar social characteristics. On the other hand, structural homophily refers to the fact that people with similar social characteristics may prefer similar physical attributes of neighborhoods. Due to these two types of homophily, residents with homogeneous characteristics will move to similar neighborhoods resulting in spatial patterns of socio-economic and demographic characteristics. This can explain spatially correlated purchasing behavior that is not created by the direct influence of one's behavior on another.

2.3. Exogenous shocks

A last cause of the existence of interdependence between customers is exogenous shocks. People of the same neighborhood may buy similar products or brands neither because they are influenced by each other nor because they have similar characteristics, but because they are subject to the same exogenous shock that exists in the neighborhood, such as promotional activities, the location of points of sales or even typical characteristics of the environment in the neighborhood.

3. Methodology

3.1. Data Description

Data is collected from one of the largest external data vendors in Belgium. This external data vendor possesses data about socio-demographics and lifestyle variables from more than 3 million respondents in Belgium. Furthermore, it provides information about automobile ownership in December 2007 of a Japanese automobile brand.

Table 1 gives an overview of all variables used throughout this study. The purpose of the proposed model is identifying respondents with a similar profile as current owners of the Japanese automobile brand, who can then be targeted using a marketing acquisition campaign. Hence, this customer identification model uses a binary variable as dependent variable, indicating whether the subject possesses the Japanese car brand. A customer acquisition model often cannot rely on transactional information because company's customer databases are typically single source in nature and do not contain information about non-customers [16]. Consequently, only a high number of socio-demographic and lifestyle predictors can be attracted from an external data vendor. The socio demographic variables contain variables that are traditionally included in a customer acquisition model. All categorical variables are split into n-1 dummies before they were included into the model. The lifestyle variables are variables created by the external data vendor indicating the interest of the respondent in a certain product category. These ratio summary variables were created based on multiple underlying questions and range from 0, if the respondent has totally no interest in the product category, to 1, if the respondent's interest is very high. Taking also these life-style variables into account should prevent that the extra value resulting from incorporating neighborhood effects is caused by the absence of other important predictors that easily could be obtained from an external data vendor.

Variable name	**Description**
Dependent variable:	
Ownership	A binary variable indicating whether the subject possesses a particular Japanese automobile brand
Independent variables:	
Socio-demographic variables:	
Age	The subject age divided into 14 age groups
Gender	The gender of the subject
Income	The income of the subject divided into 5 classes
Language	The language of the subject

Head_of_family	Whether the subject is head of the household
Pers_fam	The number persons in the household of the subject
Kid	The number of kids in the household of the subject divided into 4 age groups
Director	The subject is self-employed, a director, a manager at a puplic limited company or a manager at a private limited company
Nb_household	The number of households in the building of the subject

Lifestyle variables:
26 variables ranging from 0 to 1 indicating the interest of a subject into particular product categories: *Active sports, Cars, Cell phone, Cleaning products, Clothes, Consumer credits, Culture, Decoration, Extra insurance, Food and drinks, Grocery shopping, Holidays, Internet, Magazines, Multimedia, Multimedia equipment, Newspapers, Non-profit, No-risk investments, Omnium insurance, Risk investments, Passive sports, Pay-TV, Personal hygiene, Telephoning, Wellness*

Table 1: Model variables

Besides this data, also information about the geographical location of the respondents is needed. For this, spatial variables are used provided by the external data vendor company that divides customers into mutually exclusive neighborhoods (e.g. zip-codes). Such variables can be obtained easily and, as a result, frequently used for spatial analysis in marketing [6,9,14]. These neighborhood indicators are often constructed on multiple levels of granularity (e.g. country, district, ward, etc.). Hence, the level on which the respondents are grouped can have an influence on the predicted performance of the model. Therefore, this study will investigate a wide variety of granularity levels offered by the external data vendor. Table 2 presents the seven granularity levels examined in this study in combination with information about the number of neighborhoods at that level, the average number of respondents and the average number of owners (of a particular product) in each neighborhood. Comparing the number of owners to the total number of observations indicates that the percentage of owners is relatively small (i.e. 0.88 %).This results from the facts that, firstly, not every respondent owns a car and, secondly, there exists a lot of competition in the automobile market resulting in a wide range of automobile brands to choose from.

Granularity level	Number of neighborhoods	Average number of respondents	Average number of owners
level 1	9	349281.78	3073.00
level 2	43	73105.49	643.19
level 3	589	5337.07	46.96
level 4	3092	1016.67	8.94
level 5	6738	466.54	4.10
level 6	19272	163.11	1.44
level 7	156089	20.14	0.18

Table 2: Overview of granularity levels

Analysis based on a finer level of granularity will divide the respondents over more neighborhoods resulting in a smaller number of interdependent neighbors. At the finest level, an average of about 20 respondents is present in each neighborhood, which corresponds with an average of only 0.18 owners per neighborhood. This study will investigate which granularity level is optimal to incorporate customer interdependence using a generalized linear autologistic regression model, but also how the sample size can influence the power of these spatial variables.

3.2. Generalized Linear Autologistic Regression Model

A typical data mining technique used in CRM to solve a binary classification problem is a logistic regression model. This model is very popular in CRM because of its interpretability. Unlike other, more complex predictive techniques (e.g. neural networks), logistic regression is able to provide information about the size and direction of the effects of the independent variables [26,27].

A key assumption of this traditional model is that the behavior of one individual is independent of the behavior of another individual. Though, in reality, a customer's behavior is not only dependent of its own characteristics but is also influenced by the preferences of others. In traditional data mining techniques this interdependence is treated as nuisance in the error term. However, an autologistic regression model can be used to consider spatial autocorrelation explicitly in a predictive model for a binary variable. Originally, this model has been used in biological sciences [28-30], but recently it is also introduced in the field of marketing [10]. The generalized linear autologistic regression model in this study is a modified version of the general autologistic model introduced by Besang [31, 32]:

$$P(y = 1\ |all\ other\ values) = \frac{exp\ (\eta)}{1 + exp\ (\eta)}. \qquad (1)$$
$$Where\ \ \eta = \beta_0 + X\beta_1 + \rho WY\ .$$

In this equation a logit link function is used to adopt the regression equation to a binomial outcome variable. Whereby *Y* is an *n x 1* vector of the dependent variable; *X* is an *n x k* matrix containing the explanatory variables; the intercept is represented by β_0 and β_1 is a *k x 1* vector of regression coefficients to be estimated.

This model includes also a spatial lag effect by means of the autoregressive coefficient *ρ* to be estimated for the spatially lagged dependent variables *WY*. These spatially lagged dependent variables are constructed based on a spatial weight matrix *W*.

The weight matrix is an important element in a generalized linear autologistic regression model and can be constructed in several ways. One way of creating the spatial weight matrix is based on the continuous distance between customers. Reference [13] for example assumed that geographical influence is an inverse function of geographical distance by using the following formula:

$$w_{ij} = \frac{1}{exp\ [d(i,j)]}. \qquad (2)$$

In which *d(i,j)* represents the Euclidian distance calculated based on the latitude and longitude coordinates of the customers.

Within the field of marketing though, often a discrete spatial variable is used that divides customers into mutual exclusive neighborhoods (e.g. zip-codes) [6,9,14]. For such kind of variables the use of a contiguity matrix is more appropriate. Such matrix is constructed based on the relative positions of one customer to another. Since this study is focused on comparing discrete neighborhood variables, also a contiguity matrix will be used. This weight matrix is constructed based on an *n x n* matrix containing the elements w_{ij} indicating the interdependence between observation *i* (row) and *j* (column). Similar as in reference [13], w_{ij} will be set to one in a non-standardized weight matrix for customers living in the same neighborhood. By convention, self-influence is excluded such that diagonal elements w_{ij} equal zero. Next, this weight matrix is row-standardized using the following formula:

$$w_{ij}^{s} = \frac{w_{ij}}{\sum_j w_{ij}} . \qquad (3)$$

Hence, at a coarse granularity level the amount of neighborhoods is small resulting in a high number of interdependent relationships included in the weight matrix. Consequently, the importance of the interdependent relationships of the customers that have an influence in reality could fade away because too much interdependence is assumed. As the granularity level becomes finer, the number of non-zero elements in the weight matrix will drop. However, if the level of granularity is too fine, the number of interdependent relationships could be too small, affecting the stability of the spatial lag effect. Therefore, this study will also investigate how the sample size of the dataset could influence the optimal granularity level.

Since the correlation among customers' purchasing behavior can have several origins (e.g. word-of-mouth and homophily), it is possible that this neighborhood effect can be divided into several sub-effects, each optimally estimated at a different granularity level. Hence, this paper will apply a model that incorporates spatial autocorrelation at multiple levels of granularity using the following formula:

$$P(y = 1\ |all\ other\ values) = \frac{exp\ (\eta)}{1 + exp\ (\eta)}\ .$$
$$Where\ \eta = \beta_0 + X\beta_1 + \sum_g \rho_g W_g Y\ . \tag{4}$$

In this model a separate autoregressive coefficient is estimated for each weight matrix constructed based on a different granularity level g. This should allow the model to incorporate each variety of spatial autocorrelation using its optimal measurement level, resulting in a more accurate predictive model.

Because this study is based on a high number of observations and variables, all model parameters are obtained using a maximum pseudolikelihood (MPL) estimation. Although more advanced techniques, such as Markov chain Monte Carlo (MCMC) [33] methods have been discussed in the literature, these techniques are not implemented because they are computationally infeasible for this large database. Furthermore, Reference [34] suggests that MPL estimates should be adequate when the spatial autoregressive coefficient is relatively small. In contrast to biological sciences, this is mostly the case in the field of marketing.

The model also includes a backward selection at a significance level of 0.0001 to eliminate redundant variables that do not add extra predictive value. This should improve the comprehensibility of the model and decrease computational time and cost for scoring respondents [35].

3.3. Evaluation Criterion

In order to evaluate the predictive performance of the model, the database, containing more than 3 million observations, is randomly split into two parts. A training sample, consisting of 70% of the observations, is used to estimate the model. Afterwards, this model is validated on the remaining 30% of observations. Several evaluation criteria, such as lift or PCC (percent correctly classified), suffer from the limitation that a cutoff value needs to be chosen to discriminate predicted events from non-events. The area under the receiver operating characteristic curve (AUC) solves this limitation by taking all possible thresholds into account [36]. The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of sensitivity (i.e. the number of true positives versus the total number of events) and one minus specificity (i.e. the number of true negatives versus the total number of non-events) for all possible cutoff values used. The area under this curve can range from a lower limit of 0.5 to an upper limit of 1. The closer this value is to one, the better the general accuracy of the model.

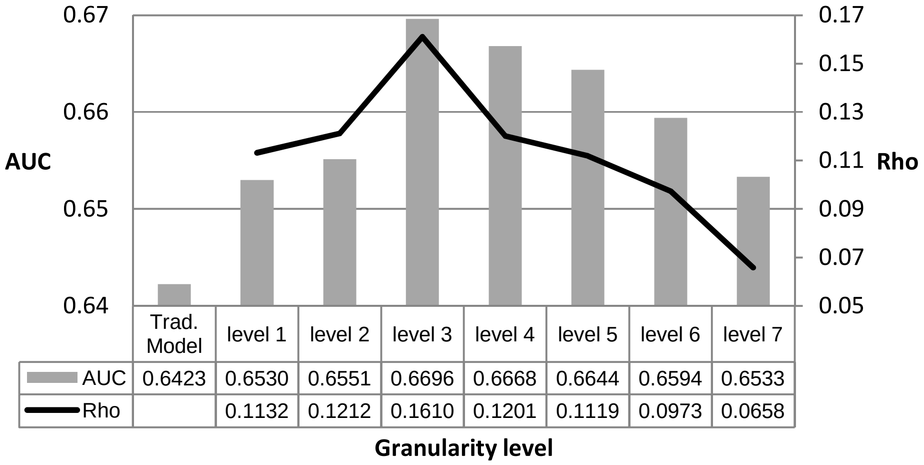
4. Results

In this chapter an overview of the results will be presented. In the first section a traditional logistic regression is compared with seven “single level” autologistic models that include spatial interdependence, each calculated based on a different level of granularity. Next, in the second section the best performing “single level” autologistic model is compared with a model that incorporates all levels of granularity simultaneously. In the last section, the effect of the sample size is examined on the predictive performance of the spatial models.

4.1. “Single level” autologistic model

In Fig.1, the traditional customer identification model and all “single level” spatial models are compared. This figure presents for each model the predictive performance on the validation sample in terms of AUC and the autoregressive coefficients estimated by the spatial models.

These spatial autoregressive coefficients are positive and significantly different from zero in all autologistic regression models. This suggests the existence of interdependence at all levels of granularity. In other words, the average correlation between automobile preferences of respondents in the same neighborhood is higher than the average correlation between automobile preferences of respondents located in different neighborhoods. Comparing the AUC indicators of the spatial models with the benchmark traditional logistic regression model using the non-parametric test of Delong et al. [37], demonstrates that incorporating these neighborhood effects significantly improves the accuracy of the acquisition model.



	Trad. Model	level 1	level 2	level 3	level 4	level 5	level 6	level 7
■ AUC	0.6423	0.6530	0.6551	0.6696	0.6668	0.6644	0.6594	0.6533
━ Rho		0.1132	0.1212	0.1610	0.1201	0.1119	0.0973	0.0658

Figure 1: Overview of the AUCs and the spatial autoregressive coefficients

However, the proportion of this predictive improvement heavily depends on the chosen granularity level. The optimal predictive performance in this study is achieved at granularity level

3. If the neighborhood level is too coarse, correlation is assumed between too many customers that do not influence each other in reality. On the other hand, a model based on a granularity level that is too fine could ignore interdependent relationships that exist in reality and affect the stability of the spatial lag effect because the number of customers in each neighborhood is too small. A similar evolution can be found in the spatial autoregressive coefficient (rho), which represents the existence of spatial interdependence in the model.

Comparing the predictive performance of a customer acquisition model that incorporates neighborhood effects at the optimal granularity level with the benchmark traditional logistic regression model illustrates that taking spatial correlation into account heavily increases the AUC by 2.73%. Although the differences between AUC can seem quite small, Reference [14] has illustrated that since such models are typically applied on a large number of prospects, even small differences in AUC can lead to large differences in terms of profitability. In other words, this improvement in predictive performance is not only statistically significant, but also economically relevant and should help marketing decision makers to improve their customer acquisition strategies.

4.2. "All levels" autologistic model

In Table 3, a comparison is made between the benchmark logistic regression model, the best performing spatial model at granularity level 3 and a model that simultaneously includes all granularity levels. This table gives an overview of all standardized parameter estimates of the socio-demographic and lifestyle variables that significantly influence automobile purchasing behavior at a 0.0001 significance level; the significant spatial autoregressive coefficients and the predictive performance of each model in terms of AUC.

Variable	Stand. est. benchmark model	Stand. est. spatial model (level 3)	Stand. est. spatial model (all levels)
Socio-demographic variables:			
Age group 18-21	-0.0548	-0.0586	-0.0592
Age group 22-25	-0.0241	-0.0256	-0.0264
Age group 31-35	-0.0292	-0.0260	-0.0268
Age group 36-40	-0.0359	-0.0345	-0.0356
Age group 61-65		0.0164	
Age group 66-70	0.0207	0.0235	0.0205
Age group 71-75	0.0165	0.0202	0.0175
Age group 76-80	0.0194	0.0230	0.0205
Is no director, self-employed earner or manager	0.0451	0.0437	0.0435
Manager at a private limited company	-0.0276	-0.0288	-0.0293
Number of persons in the household	-0.0669	-0.0628	-0.0662
Head of the household	-0.0614	-0.0547	-0.0553
Number of children younger than 5	-0.0222	-0.0232	-0.0228
Lifestyle variables:			
Cars	0.1265	0.1276	0.1262
Grocery shopping	0.1019	0.1003	0.1008
Magazines	0.0568	0.0542	0.0531
Clothes	-0.0541	-0.0633	-0.0590
Omnium insurance	-0.0439	-0.0374	-0.0355
Personal hygiene	0.0407	0.0467	0.0441
Passive sports	0.0375	0.0354	0.0380
Active sports	-0.0372	-0.0341	-0.0359
No risk investments	0.0369	0.0393	0.0397
Food and drinks	-0.0356	-0.0367	-0.0364
Cell phones	0.0299	0.0329	0.0329
Wellness	-0.0292	-0.0288	-0.0321
Consumer credit	0.0276	0.0282	0.0283
Newspapers	-0.0253	-0.0277	-0.0273
Culture	-0.0240	-0.0262	-0.0263
Telephoning	-0.0237		
Pay TV	0.0188		
Non-profit organizations		0.0201	0.0224
Spatial autoregressive coefficients (ρ):			
level 1			0.0412
level 3		0.1610	0.0935
level 4			0.0337
level 5			0.0299
level 7			0.0485
AUC:	**0.6423**	**0.6696**	**0.6783**

Table 3: Overview of the parameter estimates of the benchmark model, the spatial model at granularity level 3 and the spatial model including all granularity levels

Among the socio-demographic variables, age is a significant predictor. Older people are more likely to drive the Japanese automobile brand than younger people. Among the lifestyle variables, it is obvious that people who are more interested in cars are more likely to purchase the Japanese automobile brand. The parameter estimates of the three models do not differ a lot in size and direction. Except for one age group, i.e. age group 61-65, all the same socio-demographic variables are significant. Considering the lifestyle variables, telephoning and pay TV turn out to be only significant in the benchmark model, whereas interest in non-profit organizations is only significant in the two spatial models.

More remarkable is that the spatial autoregressive coefficient already has the strongest influence of all parameters in the spatial model at granularity level 3. This again, points to the importance of incorporating spatial correlation in customer acquisition models at the correct level of granularity.

Comparing the spatial model that includes all granularity levels with the spatial model at the optimal level proves the value of simultaneously including all granularity levels. Whereas in the first model all neighborhood effects need to be captured in one spatial autoregressive coefficient, the second model makes it possible to estimate spatial correlation at several granularity levels. As a result, the spatial autoregressive coefficients are significant at five different neighborhood levels. Interdependence between customers' purchasing behavior is still best measured at level 3, but the model is also able to capture neighborhood effects on a coarser level 1 and several finer granularity levels (i.e. level 4, 5 and 7). The spatial autoregressive coefficients at level 2 and level 6 are not significant at a 0.0001 significance level. The spatial interdependencies measured by these two spatial lag effects are already covered by other spatial variables.

Such a model is able to improve the AUC with an extra 0.87% compared to the best spatial model based on a single weight matrix which means a total improvement of 3.60% compared to a traditional CRM model. These results suggest that if the company has the resources to acquire multiple measurement levels of neighborhoods, it is advisable to simultaneously include them in a spatial CRM model in order to obtain even more accurate predictions.

4.3. Sample size effect

In an autologistic model, spatial interdependence is incorporated based on a spatial lag effect that represents the purchasing behavior of neighboring customers. However, at finer granularity levels the number of observations within such matrix can become too small, affecting the stability of the spatial influence. As a result, the sample size of the dataset can have an influence on the effect of these spatial parameters. In order to investigate this, smaller samples of the original dataset are generated. Table 4 gives an overview of the different sample sizes examined. Each sample is generated by randomly selecting a number of observations from the original dataset. In this way, 10 datasets are created for each sample size. Except for sample size “100%”, for which only the original dataset is used.

Sample size	Average number of observations	Average number of events
2%	62871	563.30
4%	125742	1094.90
6%	188613	1671.90
8%	251483	2211.00
10%	314354	2754.10
20%	628708	5550.50
40%	1257415	11075.00
60%	1886122	16612.60
80%	2514829	22102.70
100%	3143536	27657.00

Table 4: Overview of sample sizes

Similar as done for the original dataset, each of the 90 newly created samples are split into a training (70%) and a validation sample (30%). On each of these training samples, a traditional model, 7 “single level” and an “all levels” autologistic model are estimated. Next, based on the validation sample, the predictive performance of these models is calculated in terms of AUC. The average predictive performance per sample size is presented in Fig. 2. First of all, this figure clearly illustrates the value of large datasets. In general, this figure indicates for all models that the larger the sample size is, the higher the predictive performance on the validation sample. However, this effect is even larger for the spatial models than for a traditional logistic regression model. This illustrates again the importance of collecting enough data to construct stable spatial lag effects. Secondly, the larger the sample size, the more granularity level 3 emerges as optimal granularity level. When the sample size becomes smaller, the difference in predictive performance with spatial models based on coarser granularity levels becomes smaller. For sample size "2%" and "4%", the spatial models on granularity level 1 even outperform the level 3 models. In other words, the optimal granularity level tends to move to a coarser level as a result of the smaller sample size.

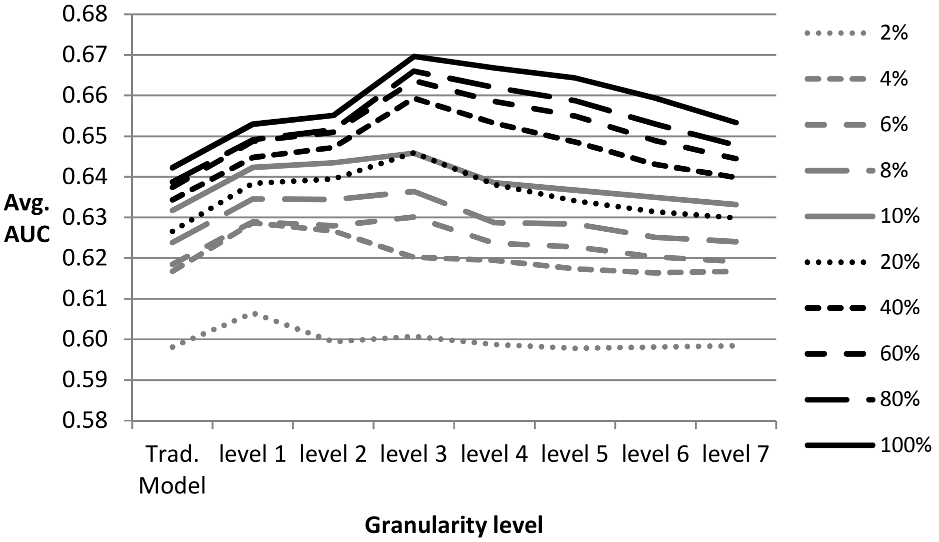


Figure 2: Overview of the average AUCs at different sample sizes

Fig. 3 explains this tendency by plotting the average spatial autoregressive coefficient of the "single level" autologistic models at several sample sizes. This figure shows that the sample size has an important effect on the spatial autoregressive coefficient (rho). In general, the spatial predictors become more important when the sample size increases. However, for the spatial autoregressive coefficient calculated on a coarse level of granularity, already a small data sample is sufficient to obtain a strong effect on the dependent variable. More specifically, for level 1 and level 2, the spatial autoregressive coefficient remains relative constant starting from sample size “6%”. From this point, the spatial lag effects are constructed based on enough neighbors to be sufficiently stable. Similarly, the spatial autoregressive coefficient at level 3 flats out starting from sample size “60%”. At this granularity level, more neighborhoods are used to incorporate spatial interdependence. As a result, more observations are needed to construct stable spatial lag effects. The spatial variables constructed on even finer levels of granularity show a very small influence in the models based on small sample sizes, but once more data is available to construct better spatial lag effects, the impact of these spatial variables is clearly improving.

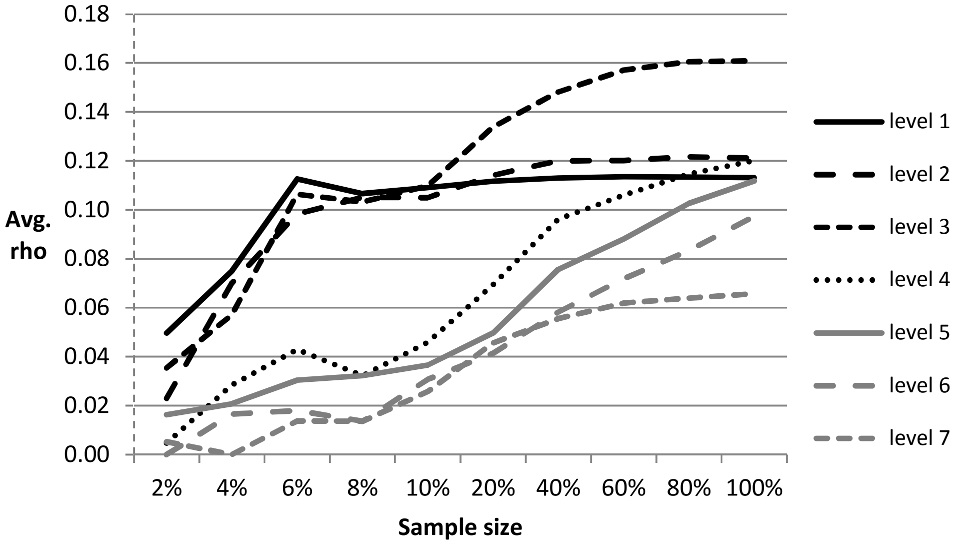


Figure 3: Overview of the average spatial autoregressive coefficients (rho) of the “single level” autologistic models at different sample sizes

Sample size	Avg. AUC best “single level” autologistic model	Avg. AUC “all levels” autologistic model	AUC difference
2%	0.6065*	0.6027	-0.0038
4%	0.6286*	0.6294	0.0008
6%	0.6301**	0.6357	0.0056
8%	0.6364**	0.6406	0.0042
10%	0.6458**	0.6508	0.0050
20%	0.6459**	0.6506	0.0047
40%	0.6594**	0.6644	0.0050
60%	0.6636**	0.6696	0.0060
80%	0.6660**	0.6730	0.0070
100%	0.6696**	0.6783	0.0087

* Based on level 1 model
** Based on level 3 model

Table 5: Comparison of the Average AUC between “single level” and “all levels” autologistic model at different sample sizes

Finally, the effect of sample size is also examined for an autologistic model that simultaneously incorporates all levels of granularity. Table 5 makes a comparison of the predictive performance between such model and the best performing “single level” autologistic model at multiple sample sizes. Again the predictive performance is expressed in terms of the average AUC over 10 randomly created datasets for each sample size. For sample size “100%”, only the original dataset is used. For sample size “2%” and “4%”, the level 1 model emerges as best performing “single level” model. Starting from sample size “6%” the data sample is large enough for the level 3 model to become superior. In the last column of Table 5 the difference between both a “single level” and “all levels” model is demonstrated. This clearly shows that the larger the data sample, the more one can benefit from the advantages of the extended autologistic model. At small sample sizes an “all levels” model is not able to outperform a “single level” model. At the smallest sample size these models perform even worse than a “single level” model. This is because on the training sample spatial variables created at finer granularity levels can become significant, but these variables have more the tendency to disturb predictions on the validation sample because they are not sufficiently stable. Once the data sample become larger, the predictive improvement, as a result of including multiple levels of granularity simultaneously, increases gradually.

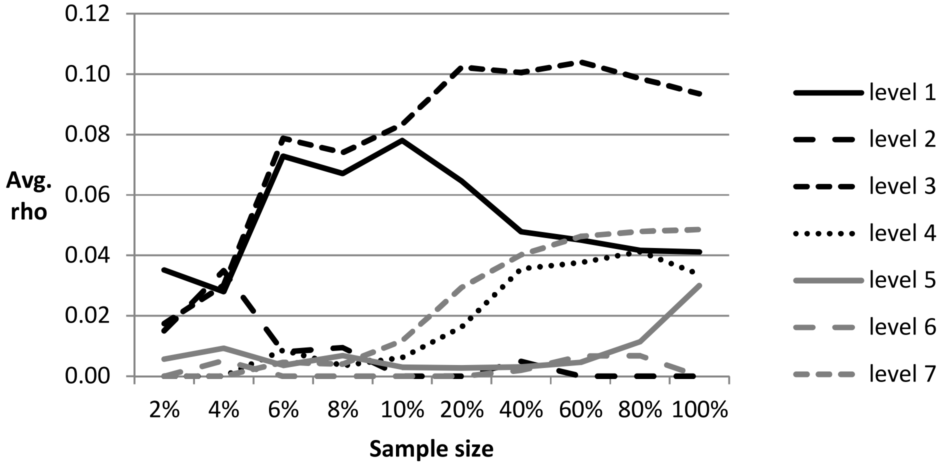


Figure 4: Overview of the average spatial autoregressive coefficients (rho) of the "all levels" autologistic models at different sample sizes

Fig 4. Explains this evolution by graphically representing the average spatial autoregressive coefficients of these extended autologistic models. This figure shows a similar trend as observed in the "single level" models. The autoregressive coefficients at a coarser level become powerful rather quickly even at small sample sizes. When more data becomes available also the spatial variables calculated on a finer granularity level are gaining importance. By this, the model is better able to distinguish several origins of spatial interdependence using multiple spatial weight matrices, resulting in an increasing improvement of predictive performance. Actually, this graph shows that once enough data is available to construct more stable spatial lag effects at a finer granularity level, some of the spatial interdependence that is firstly explained by the level 1 spatial variable can be better explained on a finer level of granularity. In contrast to Fig. 3, some spatial autoregressive coefficients remain low in the "all levels" autologistic model because the spatial interdependencies measured by these spatial variables are already covered by other spatial variables.

5. Discussion and conclusion

Traditional customer acquisition models often ignore the spatial correlation that could exist between the purchasing behaviors of neighboring customers and treats this as nuisance in the error term. Based on data of a Japanese automobile brand, this study shows that, even in a model that already includes a large number of socio-demographic and lifestyle variables typically attracted for customer acquisition, extra predictive value can still be obtained by taking spatial interdependence into account using a generalized linear autologistic regression model.

Further, this study indicates that the marketing decision maker should carefully choose the granularity level on which the neighborhoods are composed because this can have an important impact on the model's accuracy. In this research, the best predictive performance was obtained at granularity level 3. Estimations based at coarser granularity levels include too much interdependence that does not exist in reality, affecting the validity of the model. Conversely, if the level of granularity becomes too fine, the number of observations and events in each neighborhood declines, which can affect the stability of the spatial lag effect. Further, correlation could be ignored with customers that still have an influence in reality.

This study also points out that the existence of neighborhood effects can have multiple origins, such as social influences, homophily, and exogenous shocks. As a result, the underlying interdependence can be divided into multiple parts, each optimally measured on a different level of granularity. This paper shows that a model that simultaneously includes multiple granularity levels is able to outperform the best generalized linear autologistic regression model based on a single weight matrix. Hence, if the marketing decision maker has sufficient recourses it is advisable to obtain data which divides customers into neighborhoods at multiple granularity levels. This simplifies the decision to select optimal neighborhood level because this model is able to simultaneously incorporate all levels and automatically divide the existing interdependence, this causes each underling effect to be estimated based on its optimal granularity level.

In a sensitivity analysis, this study demonstrates how the sample size can influence the effect of the spatial variables. Spatial influences are included based on a spatial lag effect that incorporates the purchasing behavior of surrounding customers living in the same neighborhood. Hence, this study shows that using a finer level of granularity is only valuable when enough data is available. If not, the spatial lag effect will be calculated based on too few observations, which affects the stability of this variable. Consequently, when the data sample becomes smaller, the optimal level of granularity tends to move towards a coarser level. In addition, this also affects the use of a model that simultaneously takes multiple granularity levels into account. In order to take advantage of the fact that each origin of spatial interdependence can be measured at its optimal level, stable spatial lag effects need to be constructed even on finer levels of granularity. As a result, the difference in predictive improvement between such extended model and a "single level" autologistic model increases gradually when the data sample becomes larger.

Although this study provides interesting insights, there are still some recommendations for future research. This study is executed on a specific CRM model for a specific product. It examines the incorporation of neighborhood effects in a customer identification model that predicts automobile preferences for a Japanese automobile brand. In order to generalize the conclusions in this study, future research should verify these findings in different contexts. First of all, this highly visible and luxury good is a perfect example for which social influences and spatial interdependence can be suspected. Further research could also investigate the effect of the chosen granularity level in a context of less visible or luxury goods. Secondly, data augmentation is crucial in customer acquisition models because no transactional information is typically available, but incorporating spatial autocorrelation could also be valuable in other CRM disciplines, such as customer development or churn models. Finally, this study points out that the choice of neighborhood level can have an important influence on the model’s accuracy. This study already examined the influence of sample size on the optimal granularity level, but further research could search for other elements that might have an influence on this optimal level.

Acknowledgement

Both authors acknowledge the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

References

1. Ngai, E. W. T., Xiu, L., Chau, D. C. K.: Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Appl. 36, 2592--2602 (2009)

2. Petrison, L. A., Blattberg, R. C., Wang, P. : Database marketing past, present and future. Journal of Direct Marketing 7, 27--43 (1993)

3. Ling, R., Yen, D. C. : Customer relationship management: An analysis framework and implementation strategies. Journal of Computer Information Systems 41, 82--97 (2001)

4. Kamakura, W., Mela, C. F., Ansari, A. , Bodapati, A. ,Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B. , Verhoef, P. C. , Wedel M., Wilcox, R.: Choice models and customer relationship management. Mark. Lett. 16,279--291 (2005)

5. Baecke, P., Van den Poel, D.: Improving Purchasing Behavior Predictions by Data Augmentation with Situational Variables. Int. J. Inf. Technol. Decis. Mak. 9, 853--872 (2010)

6. Bradlow, E.T., Bronnenberg, B., Russell, G.J., Arora, N., Bell, D.R., Duvvuri, S.D., TerHofstede, F., Sismeiro, C., Thomadsen, R., Yang, S.: Spatial Models in Marketing. Mark. Lett. 16, 267--278 (2005)

7. Bronnenberg, B.J.: Spatial models in marketing research and practice. Appl. Stoch. Models. Bus. Ind. 21, 335--343 (2005)

8. Bronnenberg, B.J., Mahajan, V.: Unobserved Retailer Behavior in Multimarket Data: Joint Spatial Dependence in Market Shares and Promotional Variables. Mark. Sci. 20, 284--299 (2001)

9. Bell, D.R., Song, S.: Neighborhood effects and trail on the Internet: Evidence from online grocery retailing. QME-Quant. Mark. Econ. 5, 361--400 (2007)

10. Moon, S., Russel, G.J.: Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. Mark. Sci. 54, 71--82 (2008)

11. Grinblatt, M., Keloharju, M., Ikäheimo, S.: Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors. Rev. Econ. Stat. 90, 735--753 (2008)

12. Manchanda, P., Xie, Y., Youn, N. The Role of Targeted Communication and Contagion in Product Adoption. Mark. Sci. 27, 961--976 (2008)

13. Yang, S., Allenby, G.M.: Modeling Interdependent Customer Preferences. J. Mark. Res. 40, 282--294 (2003)

14. Steenburgh, T.J., Ainslie, A.: Massively Categorical Variables: Revealing the Information in Zip Codes, Mark. Sci. 22, 40--57 (2003)

15. Baecke, P., Van den Poel, D.: Including spatial interdependence in customer acquisition models: a cross-category comparison. Expert Syst. Appl. 39, 12105--12113 (2012)

16. Baecke, P., Van den Poel, D.: Data augmentation by predicting spending pleasure using commercially available external data. J. Intell. Inf. Syst. 36, 367--383 (2011)

17. Arndt, J.: Role of Product-Related Conversations in the Diffusion of a new Product. J. Mark. 4, 291--295 (1967)

18. Allsop, D.T., Bassett, B.R., Hoskins, J.A.: Word-of-mouth Research: Principles and Applications. J. Advert. Res. 47, 398--411 (2007)

19. Keller, E.: Unleashing the Power of Word of Mouth: Creating Brand Advocacy to Drive Growth. J. Advert. Res. 47, 448-452 (2007)

20. Carl, W.: What's all the Buzz about? Everyday Communication and the Relational Basis of Word-of-Mouth and Buzz Marketing Practices. Manag. Com. Q. 19, 601--634 (2006)

21. Bikhchandani, S., Hirshleifer, D., Welch, I.: A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. J. Polit. Econ. 100, 992--1026 (1992)

22. Chen, Y., Wang, Q.I., Xie, J.: Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning. J. Mark. Res. 48, 238--254 (2011)

23. Hartmann, W.R., Nair N., Manchanda P., Bothner M., Dodds P., Godes D., Hosanagar K., Tucker C: Modeling social interactions: identification, empirical methods and policy implications. Mark. Lett. 19, 287--304 (2008)

24. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in Social Networks. Annu. Rev. Sociol. 27, 415--444 (2001)

25. McCrea, R.: Explaining sociospatial patterns in South East Queensland, Australia: social homophily versus structural homophily. Environ. Plan. A 41, 2201--2214 (2009)

26. McCullagh, P., Nelder J.A.: Generalized linear models. Chapman & Hall, London (1989)

27. Hosmer, D.W., Lemeshow S.: Applied Logistic Regression. John Wiley & Sons, New York (2000)

28. Augustin, N. H., Mugglestone M. A., Buckland S.T.: An Autologistic Model for the Spatial Distribution of wildlife. J. Appl. Ecol. 33, 339--347 (1996)

29. Hoeting J.A., Leecaster M., Bowden D.: An Improved Model for Spatially Correlated Binary Responses. J. Agric. Biol. Environ. Stat. 5, 102--114 (2000)

30. He, F., Zhou, J., Zhu, H.: Autologistic Regression Model for the Distribution of Vegetation. J. Agric. Biol. Environ. Stat. 8, 205--222 (2003)

31. Besang, J.: Spatial Interaction and the Statistical Analysis of Lattice Systems. J. Roy. Statist. Soc. Ser. B (Methodological) 36, 192--236 (1974)

32. Besang, J.: Statistical Analysis of non-lattice data. The Statistican 24, 179--195 (1975)

33. Huffer, F.W., Wu, H.: Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species. Biometrics 54, 509--524 (1998)

34. Wu, H.,Huffer, F.W.: Modelling the distribution of plant species using the autologistic regression model. Environ. Ecol. Stat. 4, 49--64 (1997)

35. Kim, Y.S.: Toward a successful CRM: Variable selection, sampling, and ensemble. Decis. Support Syst.41, 542--553 (2006)

36. Hanley, J.H., McNeil B.J.: The meaning and use of area under a receiver operating characteristic (ROC) curve. Radiology 143, 29--36 (1982)

37. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or morecorrelated receiver operating characteristic curves: a nonparametric approach. Biometrics. 44, 837--845 (1988)

CHAPTER V:

INCLUDING SPATIAL INTERDEPENDENCE IN CUSTOMER ACQUISITION MODELS: A CROSS-CATEGORY COMPARISON

This chapter is based upon Baecke, P. & Van den Poel, D. (2012). Including spatial interdependence in customer acquisition models: a cross-category comparison, *Expert Systems with Applications, 39*, 12105 – 12113.

CHAPTER V:
INCLUDING SPATIAL INTERDEPENDENCE IN CUSTOMER ACQUISITION MODELS: A CROSS-CATEGORY COMPARISON

Abstract

Within analytical customer relationship management (CRM), customer acquisition models suffer the most from a lack of data quality because the information of potential customers is mostly limited to socio-demographic and lifestyle variables obtained from external data vendors. Particularly in this situation, taking advantage of the spatial correlation between customers can improve the predictive performance of these models. This study compares an autoregressive and hierarchical technique that both are able to incorporate spatial information in a model that can be applied to a large dataset, which is typical for CRM. Predictive performances of these models are compared in an application that identifies potential new customers for 25 products and brands. The results show that when a discrete spatial variable is used to group customers into mutually exclusive neighborhoods, a multilevel model performs at least as well as, and for a large number of durable goods even significantly better than a more often used autologistic model. Further, this application provides interesting insights for marketing decision makers. It indicates that especially for publicly consumed durable goods neighborhood effects can be identified. However, for more exclusive brands, incorporating spatial information will not always result in major predictive improvements. For these luxury products, the high spatial interdependence is mainly caused by homophily in which the spatial variable is a substitute for absent socio-demographic and lifestyle variables. As a result, these neighborhood variables lose a lot of predictive value on top of a traditional acquisition model that typically is based on such non-transactional variables.

1. Introduction

As markets become increasingly saturated and highly competitive, companies have shifted their marketing strategies from transactional marketing to relationship marketing (Coussement, Benoit, & Van den Poel, 2010; Pai & Tu, 2011). In other words, companies are more focused on the acquisition of valuable customers, the development of these customers in order to make them even more profitable and the creation of a long-term relationships in order to improve customer loyalty and retention (Kamakura et al., 2005). This is also reflected in an explosion of interest in customer relationship management (CRM) by both academics and business practitioners (Ngai et al., 2009). Due to the information revolution and the drop in costs of data warehousing, many companies have collected a vast amount of socio-demographic and transactional data of their customers. In addition, computer power is increasing rapidly and data mining techniques are used to exploit this data in an optimal manner (Hosseini, Maleki, & Gholamian, 2010; Kamakura et al., 2005). This has resulted in the development of a wide range of software tools which enable companies to transform the collected data into useful information for marketing decision makers.

As a high quality database is the foundation of effective and efficient CRM, companies should invest in augmenting their databases with extra valuable variables (Baecke & Van den Poel, 2011). In this context, several studies have proven that incorporating information about the geographic proximity between customers can be valuable in marketing (Bradlow, Russell, & Bell, 2005; Bronnenberg, 2005). This information can often be obtained at a relatively low cost and could significantly improve the performance of a CRM model. Traditional CRM models assume that customers' decisions are unrelated to each other and only depend on the characteristics of the particular customer, whereas in reality, preferences are often also influenced by the purchasing behavior of other customers and their recommendations (Arndt, 1967). Besides this, the geographical location can act as a proxy for socio-demographic information because agents with similar characteristics and tastes have the tendency to group together (Mcpherson, Smith-lovin, &

Cook, 2001). As a result of this principle, called homophily, customers within the same neighborhood are often more homogeneous in terms of socio-demographic characteristics.

Although several studies have proven the existence of spatial interdependence between the purchasing behaviors of customers (Bell & Song, 2007; Bradlow et al., 2005; Bronnenberg, 2005; Grinblatt, Keloharju, & Ika, 2008), the incorporation of spatial information in a predictive CRM context is limited. To the best of our knowledge, only two studies have incorporated spatial interdependence in order to improve customer identification, each using a different predictive technique. On the one hand, Yang & Allenby (2003) used an autoregressive approach to incorporate both geographic and demographic proximity between customers in a CRM model that predicts customers' preference for Japanese-made cars. That study indicated that geographic reference groups still have a larger impact than demographic reference groups. On the other hand, Steenburgh, Ainslie, & Hans (2003) used a hierarchical model to include a massively categorical variable, such as zip-codes, in order to improve the acquisition of new students at a private university. However, also these two studies have some limitations. Firstly, until now, both techniques have never been compared in terms of predictive performance which makes it difficult for a marketing decision maker to choose the most appropriate technique. Secondly, due to the complexity of the spatial models, both studies are based on a small number of observations and predictive variables, which does not match with current CRM applications. Thirdly, these studies were only based on one product or one university. Therefore, no real conclusion can be drawn about the applicability of these models on other product categories.

This paper contributes to these previous studies by investigating, using both an autoregressive and a hierarchical approach, how the incorporation of spatial interdependence can improve a CRM model. More specifically, this study will try to improve traditional customer acquisition models across multiple brands and products. From all CRM fields, it is often most difficult to obtain good predictive results in the case of customer acquisition. This is because obtaining information from potential customers is not straightforward (Thorleuchter, Van den Poel, & Prinzie, 2012). As a result, in order to identify possible prospects, acquisition models are often estimated only based on a limited number of variables obtained from external data vendors (Baecke & Van den Poel, 2011). Especially in such a context where the availability of data is limited, incorporating neighborhood effects can be very valuable.

This study will try to predict whether or not a respondent has bought a particular brand or product. Next, these probabilities can then be estimated on a pool of potential new customers in order to determine which of them has the highest chance to reply. Only addressing the customers with a high probability to purchase can already significantly improve the accuracy of a response model in direct marketing (W.-C. Chen, Hsu, & Hsu, 2011). Consequently, a better performing customer acquisition model can have a significant influence on a company's profit. Whereas a well-targeted mail can increase profits, an irrelevant mail will not only increase the marketing cost, but can also damage the image of a company on the long term (Kim, Lee, & Cho, 2008).

Besides, comparing two spatial techniques across multiple products and brands, another contribution of this study is the quality and quantity of the data. Table 1 illustrates that compared to previous literature this paper is based on a larger and more realistic data sample. This is necessary since this study wants to investigate the added value of spatial information on top of data traditionally used for customer identification. Hence, if only a small number of predictive variables were included, spatial information could easily become a significant predictor because it could act as a proxy for important missing variables.

Study	Spatial technique	Dependent variable	Number of observations in training sample	Number of observations in validation sample	Number of zip-codes	Number of non-spatial variables
Yang & Allenby (2003)	Hierarchical	Japanese car preference	666	191	122	6
Steenburgh et al. (2003)	Autoregressive	Enrollment of students	37,551	34,179	7279	9
This study	Hierarchical & Autoregressive	Purchasing behavior of 25 products and brands	between 237,114 and 2,200,361	between 101,621 and 943,013	589	35

Table 1: Comparison of data information between previous studies and this study

Furthermore, the application in which the effect of including spatial interdependence is compared across multiple products and brands can deliver interesting insights for a marketing decision maker. Currently, most research on spatial interdependence has been devoted to publicly consumed durable goods, such as automobiles (e.g. Grinblatt et al., 2008; Yang & Allenby, 2003). This is because these highly visible products are more likely to be subject to social influence (Bearden & Etzel, 1982). However, until now, almost no attention has been paid to the existence of neighborhood effects in less visible or less involving product categories. Therefore, besides applying spatial models on publicly consumed durable goods, this paper will also focus on privately consumed durable goods and consumer packaged goods.

The remainder of this paper is organized as follows. Section 2 will give an overview of all products and brands that will be examined in this study. In Section 3 the methodology is presented in which the two predictive models and the evaluation criteria are described. The results are reported in Section 4 and Section 5 provides a discussion of these results in combination with a conclusion.

2. Data Description and Product Categories

This paper is based on data collected from one of the largest external data vendors in Belgium. Multiple socio-demographic and lifestyle variables are used as predictors to identify customers with a preference for a particular product or brand. An overview and description of these variables can be found in Table 2.

Variable name	**Description**
Socio-demographic variables:	
Age	The subject age divided over 14 age groups
Gender	The gender of the subject
Income	The income of the subject divided over 5 classes
Language	The language of the subject
Head_of_family	Whether the subject is head of the household
Pers_fam	The number persons in the household of the subject
Kids	The number of kids in the household of the subject divided over 4 age groups
Director	The subject is a self_employed earner, a director, a manager at a puplic limited company or a manager at a private limited company
Nb_household	The number of households in the building of the subject
Lifestyle variables:	
26 variables ranging from 0 to 1 indicating the interest of a subject into particular product categories: *Active sports, Cars, Cell phone, Cleaning products, Clothes, Consumer credits, Culture, Decoration, Extra insurance, Food and drinks, Grocery shopping, Holidays, Internet, Magazines, Multimedia, Multimedia equipment, Newspapers, Non-profit, No-risk investments, Omnium insurance, Risk investments, Passive sports, Pay-TV, Personal hygiene, Telephoning, Wellness*	

Table 2: Overview of independent variables

Next to the independent variables, also a discrete zip code variable is used to group customers into 589 mutually exclusive neighborhoods. Similar to the papers of Yang & Allenby (2003) and Steenburgh et al. (2003), spatial interdependence is assumed between customers living in the same neighborhood.

		No. obs.	**No. events**
Public Durable Goods			
Automobiles	*Ford*	3143374	118192
	Toyota	3143374	85711
	Mercedes	3143374	57518
	Fiat	3143374	30759
	Volvo	3143374	26134
Clothes	*C&A*	617431	243297
	E5 Mode	617431	140613
	Zara	617431	100577
	Scapa	617431	44269
	Mango	617431	34856
Private Durable Goods			
Microwave		1348662	850068
Dish washing machine		1800293	690514
Surround system		954275	589288
Refrigerator with freezer		571372	344221
Espresso Machine		786511	121062
Consumer Packaged Goods			
Sodas	*Coca-Cola*	338735	114032
	Fanta	338735	61520
	Ice Tea	338735	54583
	Sprite	338735	41870
	Aquarius	338735	25570
Shampoos	*Dove*	342454	63626
	Elseve	342454	61845
	Fructis	342454	47003
	Pantene	342454	42560
	Head & Shoulders	342454	39237

Table 3: Overview of examined products and brands

This paper gives an overview for which products and brands spatial interdependence can be observed and investigates whether taking the spatial structure of the data into account can improve CRM predictions for customer acquisition. Table 3 presents all products and brands examined in this study, divided into three main groups, namely public durable goods, private durable goods and consumer packaged goods. As shown in the last two columns of Table 3, which represent the number of observations and the number of events of each dependent variable, this study is based on a very large data sample.

In general, research on spatial interdependence and social influence is typically carried out on durable goods, such as automobiles (e.g. Grinblatt et al., 2008; Yang & Allenby, 2003). For these products, neighborhood effects are more likely to be identified because they are purchased infrequently and relative expensive, resulting in a higher involvement of the customer. Besides involvement, also the visibility of the product could have an impact on the existence of interdependence between customers' purchasing decisions (Bearden & Etzel, 1982). Products for which the consumption is very visible will be more subject to reference group influence than privately consumed products. Therefore, durable goods are split into a publicly consumed and a privately consumed category. In the publicly consumed category five automobile brands, each brand originally coming from a different country, and five large clothing brands are examined. However in the privately consumed category, the focus will be on the purchase of five products, irrespective of the brand. This is based on Bearden & Etzel (1982) who illustrated that for publicly consumed durable goods, reference group influence mainly affects the brand choice decision, whereas for privately consumed goods the product choice decision will be mostly influenced. In each of the two durable goods categories a range of both luxury (e.g. “Mercedes”, “Volvo”, “Scapa”, “Espresso Machine”) and less luxury (e.g. “Toyota”, “C&A”, “Refrigerator with freezer”) products and brands are included, because luxuary could also have an impact on the amount of reference group influence.

Besides examining durable goods, this study will also explore the effect of incorporating spatial interdependence to identify customers of consumer packaged goods (CPGs). CPGs are typically low-involvement products with very low risk associated to the purchase. As a result, investigating the existence of spatial interdependence for these products has been ignored by literature for a long time. Only recently, two studies have discovered that during the purchase of CPGs also interdependence can exist. Kuenzel & Musters (2007) showed for low involvement products that some specific reference groups, such as close family or friends, can influence each other's purchasing behavior. Although no influence was discovered by neighbors, this study will verify this based on real behavioral data instead of questionnaires. Also Du & Kamakura (2011) detected that customers who purchased a newly introduced CPG can influence the adoption decision of neighboring customers. Although these contagion effects were mostly temporally measured during the introduction of a new CPG, this paper will investigate whether neighborhood effects can also be detected for more established CPG brands. Since these products are frequently bought by everyone, almost no differentiation would be measured in terms of purchasing behavior of the product itself. Therefore, in this category the focus will also be on brand-choice influences. Hence, ten CPG brands are included in this research divided over two product categories (i.e. sodas and shampoos).

For each of the products and brands in Table 3, this study will investigate, based on two modeling techniques, whether neighborhood effects can be observed and whether these discovered effects are strong enough to improve a traditional customer acquisition model.

3. Methodology

As previously mentioned, the purpose of an acquisition model is to predict whether or not a respondent has bought a particular brand or product. This binary classification problem is often solved in CRM by means of a logistic regression model, which will be used as benchmark model. This generalized linear model uses a logit link function to adopt ordinary least squares regression to a response variable with dichotomous outcomes (McCullagh & Nelder, 1989). The equation of this well-known model can be formulated as follows:

$$P_i(y = 1 \mid all\ other\ variables) = \frac{exp\ (\eta_i)}{1 + exp\ (\eta_i)} \tag{1}$$

$$\eta_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki}$$

whereby P_i represents the a posteriori probability that customer *i* is a buyer of a certain product; β_0 is the intercept; X_{ki} represents the independent variable k of customer *i; n* is the number of independent variables and β_k are the parameters that need to be estimated.

Several advantages have made this model a very popular technique in CRM. Unlike more complex predictive techniques, this model is easily interpretable for managers. It provides information about the size and direction of the effect of each predictor (Hosmer & Lemeshow, 2000). Further, due to its popularity, this model is widely available in many statistical packages, providing quick and robust results (Neslin et al., 2006).

Despite these advantages, an important assumption of this traditional model is that customers are assumed to act independently of other individuals. However, in reality, a customers' behavior is often influenced by the behavior and recommendation of others. Several authors already recognized that agents who are situated geographically close to each other have a higher correlating behavior (Bradlow et al., 2005; Bronnenberg, 2005). As a result, instead of treating this as nuisance in the error term, including this interdependence could improve CRM prediction.

For this end, various techniques are discussed in the literature. In most studies a spatial autoregressive model is used to capture spatial interdependence (Bell & Song, 2007; Bronnenberg & Mahajan, 2001; Yang & Allenby, 2003). Such models create a spatial weight matrix to include the behavior of surrounding agents to assist in predicting the behavior of a particular customer. However, when a spatial variable is used that divides customers into mutually exclusive neighborhoods, such as zip codes, also a hierarchical model can incorporate spatial interdependence (Steenburgh et al., 2003).

This paper will focus on two models, closely related to the models used in the research previously described, namely an autologistic model and a multilevel model. By means of both models this study will examine for multiple brands and products whether neighborhood effects can be observed. Next, the predictive improvement of these models with respect to a traditional model will be calculated. In the next two sections, the methodology of both models will be discussed.

3.1. Autologistic Model

Autologistic models have been frequently used to model the distribution of animal and plant species (Augustin, Mugglestone, & Buckland, 1996; He, Zhou, & Zhu, 2003). However, recently, the advantages of these models have also been recognized in the field of marketing (Moon & Russell, 2008). The autologistic model can be defined by means of the following equation (Besag, 1974, 1975):

(2)

$$P_i(y = 1 \,|all\ other\ variables) = \frac{exp\ (\eta_i)}{1 + exp\ (\eta_i)}$$

$$\eta_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} + \rho \frac{\sum_{i \neq j} w_{ij} Y_j}{w_{ij}}$$

This equation is similar to a logistic regression model, but a spatial lag term is included that incorporates spatial interdependency. This spatial lag term is constructed based on an autoregressive coefficient ρ to be estimated for the spatially lagged dependent variable. This spatially lagged dependent variable is calculated using a weight matrix, which contains a one for customers living in the same neighborhood and a zero for every customer combination that lives in different neighborhoods (Anselin, 1988). By convention, self-influence is excluded such that diagonal elements equal zero. Next, this weight matrix is row standardized such that all row elements sum to one and multiplied with a vector containing the observed outcome variables. As such, the predicted behavior of a customer does not only depend on the customers' own characteristics but is also assisted by the behavior of neighboring customers.

3.2. Multilevel Model

Another approach to include neighborhood effects in a binary predictive CRM model is by applying a multilevel model, also called a generalized linear mixed model (Breslow & Clayton, 1993; Wolfinger & O'Connell, 1993). This model does not include a spatial lag effect. Instead, it makes use of the hierarchical structure of the spatial data to incorporate interdependence of customers. Spatial models that specify the weight matrix as in Equation (2) are based on 'Interaction Among Places' and state that objects that are close to each other are more related than distant objects, whereas multilevel models are related to 'Place Similarity' where the focus is more on hierarchy than on proximity (Anselin, 2002; Miller, 2004). In other words, these multilevel models state that objects in the same region are more related than objects in different regions. As a result, this model is only applicable when spatial data is used that divides customers into mutually exclusive neighborhoods (e.g. zip codes). Multilevel models are widely used in social sciences (Courgeau & Baccaini, 1998; Lee & Bryk, 1989), however in marketing, only Steenburgh et al. (2003) used such model to include neighborhood effects during the acquisition process of students for a private university. Assuming that data is available from *J* neighborhoods with a different number of customers n_j for each neighborhood, the complete formula of a multilevel model can be defined as follows (Hox, 2002):

(3)

$$P_i(y = 1 \,|all\ other\ variables) = \frac{exp\ (\eta_i)}{1 + exp\ (\eta_i)}$$

$$\eta_i = \beta_{0j} + \sum_{k=1}^{n} \beta_{kj} X_{ki}$$

This formula is related to a traditional logistic regression model, but it allows the intercept and slope coefficients, β_{0j} and β_{kj}, to vary across groups. These coefficients, often called random coefficients, have a distribution with a certain mean and variance that can be explained by *l* independent variables at the highest level Z_j, as follows:

(4)

$$\beta_{0j} = \gamma_{00} + \sum_{\mathrm{m}=1}^{\mathrm{l}} \gamma_{0\mathrm{m}} Z_{mj} + u_{0j}$$

and

$$\beta_{\mathrm{k}j} = \gamma_{\mathrm{k}0} + \sum_{\mathrm{m}=1}^{\mathrm{l}} \gamma_{\mathrm{km}} Z_{mj} + u_{1j}$$

The u-terms u_{0j} and u_{1j} represent the random residual errors at the highest level and are assumed to be independent from the residual errors e_{ij} at the lowest level and normally distributed with a mean of zero and a variance of $\sigma_{u_0}^2$ and $\sigma_{u_1}^2$ respectively. Since in this model errors are not assumed to correlate, a simple diagonal covariance matrix is used which models a different variance component for each random effect.

Because this model is used in a predictive context, containing a large amount of predictive variables, it is impossible to allow all slope coefficients to vary across groups. Certainly in combination with a large number of neighborhoods the model would become too complex, which may result in overfitting. Therefore, this model is simplified to a random intercept model, which can be written as (Baecke & Van Den Poel, 2010):

(5)

$$P_i(y = 1 \mid all\ other\ variables) = \frac{exp\ (\eta_i)}{1 + exp\ (\eta_i)}$$

$$\eta_i = \beta_{0j} + \sum_{k=1}^{n} \beta_k X_{ki}$$

where

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

In contrast to an autoregressive model in which a spatial lag effect is added, this model incorporates interdependence between the purchasing behaviors of customers in the same neighborhood by varying the intercepts for each neighborhood. As a result, customers living in the same neighborhood have a higher probability to show a similar purchasing behavior than customers living in different neighborhoods.

3.3. Evaluation Criteria

In order to evaluate the predictive performance, for each product or brand, the database is randomly split into a training and validation sample. The training sample, containing 70% of the observations, is used to estimate the parameter estimates. Afterwards, each model is validated on the remaining 30% of observations. The predictive performance of each model will be expressed in terms of the area under the receiver operating characteristic curve (AUC), which is graphically presented by a two-dimensional representation of sensitivity (i.e. the true positive rate) and 1-specificity (i.e. the false positive rate) (Huang & Ling, 2005). Mathematically, AUC can be calculated using the following formula (Hand & Till, 2001):

(6)

$$\hat{A} = \frac{S_0 - n_0(n_0 + 1)/2}{n_{0\,n_1}}$$

Whereby n_0 and n_1 are the number of observations in the dataset belonging to respectively class 0 and class 1 and S_0 is the sum of the class 0 test points. This calculates the probability that a randomly chosen positive instance is correctly ranked higher than a randomly selected negative instance (Hanley & Mcneil, 1982). This probability will be close to 0.5 if predictions are random and close to 1 for perfect predictions.

An important advantage of AUC compared to other performance criteria, such as the percent correctly classified (PCC), is its independence of the chosen cut-off. The PCC gives the performance at only one cut-off level on which instances are predicted to be in class 0 or class 1, whereas the AUC gives an overall value based on all threshold values. Furthermore, Huang & Ling (2005) claimed that in general, AUC is statistically more consistent and more discriminating than accuracy.

4. Results

Before investigating whether predictive improvement can be achieved by including neighborhood effects in a CRM model, first the individual predictive effect of spatial interdependence will be examined. Therefore, an empty model is estimated without any independent variables. In other words, the autoregressive model and the multilevel model will only make use of respectively the spatial weight matrix and the hierarchical structure of the data in order to classify customers into buyers and non-buyers. This should give an indication of the amount of neighborhood effects that exists for each product or brand. In Figure 1 and 2, the predictive performance of both empty models is presented for each product and brand, divided over three product categories.

In a first step, the difference in predictive performance between an empty autologistic model and an empty multilevel model is compared by means of the non-parametric test of DeLong, DeLong, & Clarke-Pearson (1988). Since the number of observations is quite high a strict significance level of 0.001 is applied. This test indicates that for an empty model, the AUCs of both techniques do not significantly differ from each other. In other words, both models are equivalently able to measure the existence of spatial interdependence across all products and brands examined in this study.

Secondly, these figures illustrate that the existence of neighborhood effects depends on both the involvement and the visibility of the product. For public durable goods, a significant amount of customers' purchasing behavior can already be predicted by taking only the interdependent behavior of customers into account. Clearly, this is less for privately consumed durable goods and the lowest for consumer packaged goods. Next to involvement and visibility, the exclusivity of the product or brand seems to have also an influence, although to a lesser extent, on the existence of neighborhood effects. This can be derived from the relative high predictive performance of the models that predict the purchase of "Scapa", "Mercedes" and "Volvo", which are more luxury brands, compared to the other brands in their category. Also in the private durable goods category, a more luxury product such as an "Espresso machine" is ranked higher than necessities, such as a "Refrigerator with freezer".

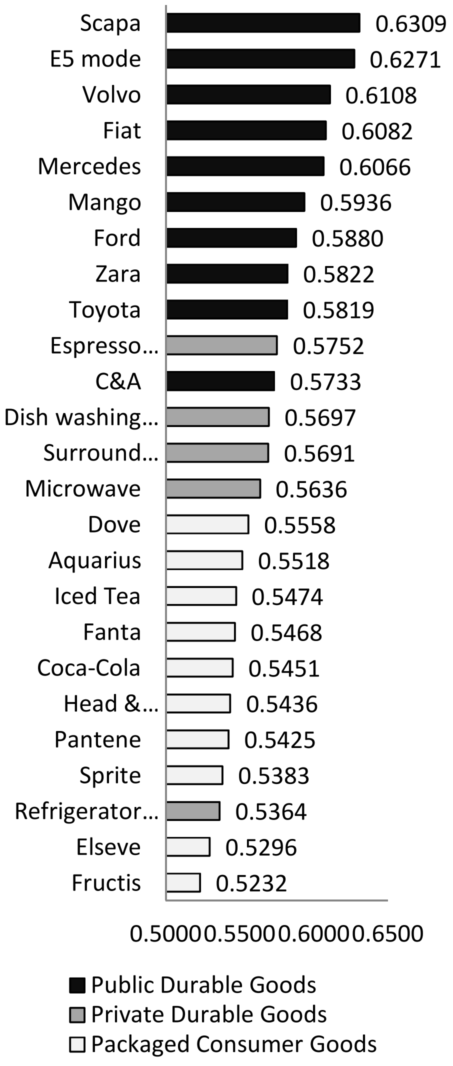


Figure 1: AUCs of an empty autologistic model

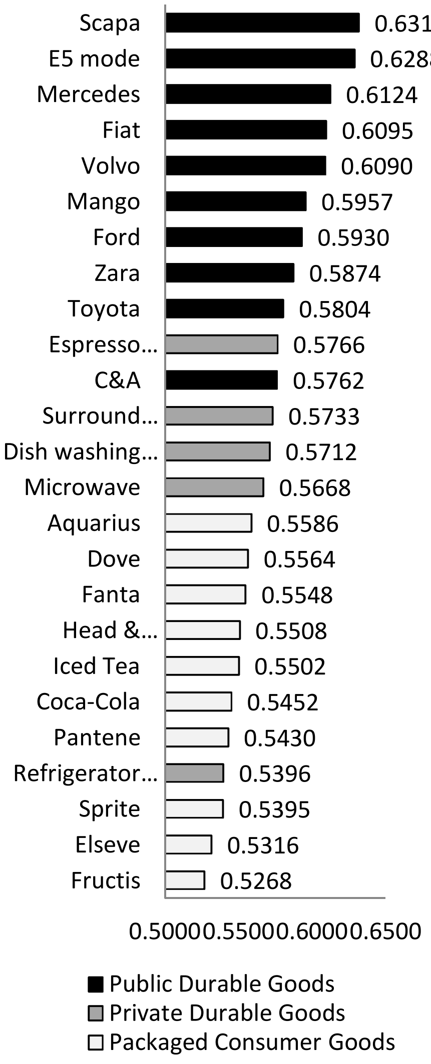


Figure 2: AUCs of an empty multilevel model

		Benchmark Model	Autologistic Model[1]	Multilevel Model[2]
Public Durable Goods				
Automobiles	*Ford*	0.6350	0.6566	0.6568
	Toyota	0.6387	0.6577	0.6582
	Mercedes	0.7399	0.7439	0.7448*
	Fiat	0.6482	0.6656	0.6674*
	Volvo	0.6976	0.7041	0.7054
Clothes	*C&A*	0.6755	0.6894	0.6922*
	E5 Mode	0.6921	0.7125	0.7131*
	Zara	0.7800	0.7885	0.7893*
	Scapa	0.8194	0.8227	0.8242*
	Mango	0.8050	0.8120	0.8117
Private Durable Goods				
Microwave		0.6993	0.7024	0.7029*
Dish washing machine		0.7220	0.7247	0.7256*
Surround system		0.7144	0.7160	0.7167*
Refrigerator with freezer		0.5947	0.5982	0.5984
Espresso Machine		0.6577	0.6624	0.6634*
Consumer Packaged Goods				
Sodas	*Coca-Cola*	0.6230	0.6240	0.6244
	Fanta	0.6882	0.6901	0.6902
	Ice Tea	0.7210	0.7227	0.7234
	Sprite	0.6958	0.6978	0.6980
	Aquarius	0.7459	0.7484	0.7493*
Shampoos	*Dove*	0.6403	0.6422	0.6423
	Elseve	0.6342	0.6364	0.6371
	Fructis	0.6732	0.6752	0.6747
	Pantene	0.6472	0.6493	0.6498
	Head & Shoulders	0.6531	0.6557	0.6556

[1] All AUCs of the autologistic model differ significantly from the benchmark model on a 0.001 significance level

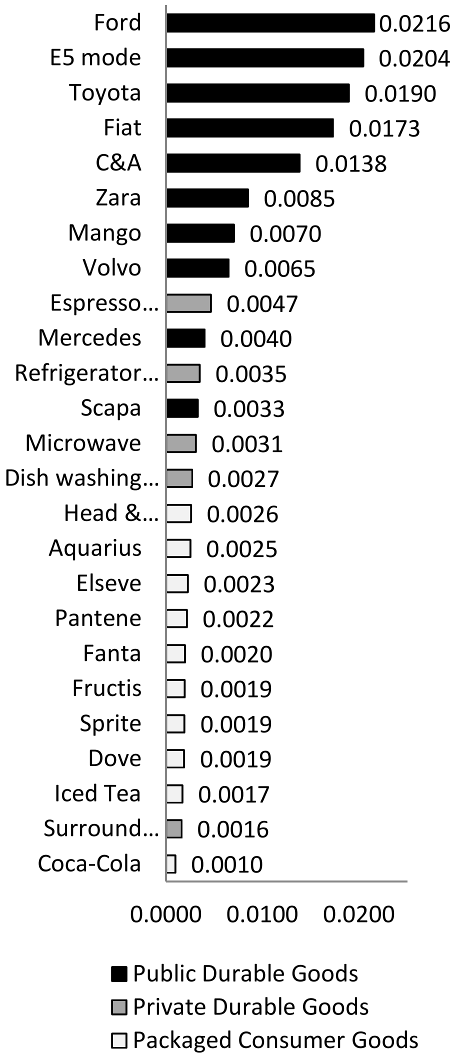
[2] All AUCs of the multilevel model differ significantly from the benchmark model on a 0.001 significance level

* Significant difference between autologistc and multilevel model on a 0.001 significance level

Table 4: Overview of the predictive performance in terms of AUC

After examining neighborhood effects individually, Table 4 demonstrates how these effects can give extra value to a customer acquisition model. This table compares for each product and brand the predictive performance in terms of AUC on the validation sample of a traditional logistic regression model, used as benchmark model, with an autologistic model and a multilevel model in which neighborhood effects are incorporated. In a comparison of the predictive performance of the models based on the non-parametric test of DeLong et al. (1988) using a 0.001 confidence interval, Table 4 shows that for all products and brands both spatial models perform significantly better than a traditional logistic regression model. This means that not only for public durable goods, but also for privately consumed durables and consumer packaged goods a significant improvement can be observed. When comparing both spatial models with each other, the results deviate from the comparison based on the empty spatial models. Although, the predictive performance between both spatial techniques is statistically equal for some product and brands, the non-parametric test of DeLong et al. (1988) indicates that in 11 of the 25 cases the multilevel model significantly outperforms the autologistic model. Especially when the purchasing behavior of durable goods is modeled, the use of a multilevel model is preferred. Since the purchases of these goods are more influenced by neighborhood effects, the way how these influences are included on top of traditional variables will have a larger impact on the total predictive performance. Hence, for these durable goods the multilevel model is superior in even 10 out of the 15 cases.

The improvement of each model is graphically represented in Figure 3 and Figure 4. In general, these figures follow the same trend as Figure 1 and 2 in such a way that also in terms of predictive improvement including neighborhood effects is most beneficial for public durable goods. Although, very remarkable is that within this product category, the most exclusive brands (i.e. “Scapa”, “Mercedes” and “Volvo”) are not able to benefit as much as the other brands while these luxuries experience the most spatial interdependence (see Figure 1 and 2).



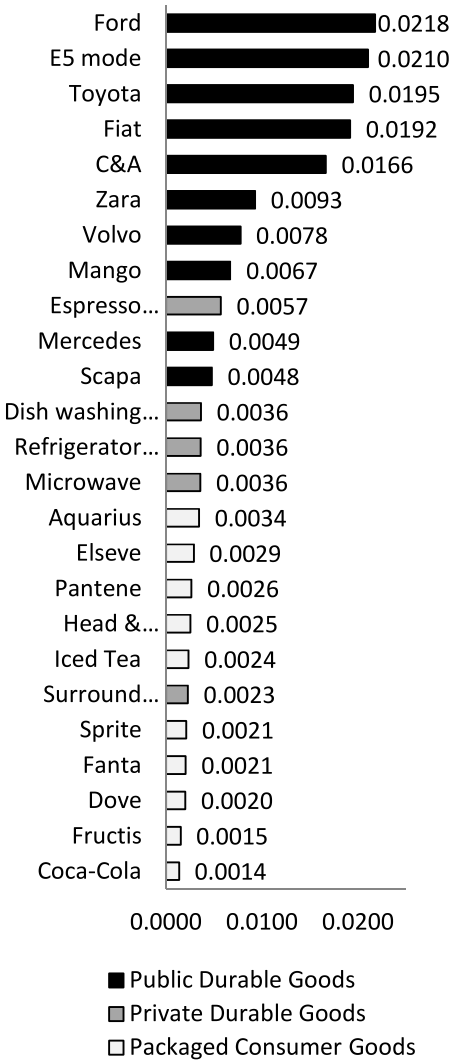


Figure 3: Predictive improvement of an autologistic model

Figure 4: Predictive improvement of a multilevel model

These luxury brands are mostly bought by a smaller, more specific group of customers. As a result, prospects can already be better identified using only socio-demographic and lifestyle variables. This is demonstrated by the high predictive performance based on only traditional variables in Table 4. In other words, the high spatial interdependence measured for these luxury brands is mainly caused by homophily in which the neighborhood variable is a substitute for the absent socio-demographic and lifestyle variables. This is also proven by table 5, in which for both predictive models the spatial parameters are compared between an empty model and a full model that includes also socio-demographic and lifestyle variables. More particular, for an autologistic model the impact of spatial interdepence is measured through the standardized spatial autoregressive coefficient, while in a multilevel model this is measured through the intercept variance estimate. All the spatial parameters in this table are significantly different from zero on a 0.001 significance level.

		Autoregressive coefficient of autologistic model		**Intercept Variance of multilevel model**	
		Empty model	**Full model**	**Empty model**	**Full model**
Public Durable Goods					
Automobiles	*Ford*	0.1558	0.1528	0.1429	0.1259
	Toyota	0.1471	0.1436	0.1211	0.1267
	Mercedes	0.1944	0.1263	0.1840	0.0412
	Fiat	0.1863	0.1678	0.1840	0.1352
	Volvo	0.1973	0.1343	0.2147	0.0713
Clothes	*C&A*	0.1535	0.1547	0.0835	0.0974
	E5 Mode	0.2532	0.2413	0.1865	0.1286
	Zara	0.1638	0.1701	0.1113	0.1136
	Scapa	0.2629	0.1895	0.2553	0.1050
	Mango	0.1770	0.1744	0.1409	0.1119
Private Durable Goods					
Microwave		0.1276	0.1146	0.0597	0.0259
Dish washing machine		0.1408	0.0877	0.0622	0.0311
Surround system		0.1443	0.1023	0.0814	0.0246
Refrigerator with freezer		0.0748	0.0553	0.0246	0.0142
Espresso Machine		0.1470	0.1076	0.0921	0.0407

Consumer Packaged Goods					
Sodas	*Coca-Cola*	0.0909	0.0605	0.0372	0.0135
	Fanta	0.0918	0.0619	0.0444	0.0202
	Ice Tea	0.0966	0.0687	0.0503	0.0293
	Sprite	0.0772	0.0636	0.0362	0.0286
	Aquarius	0.1008	0.0820	0.0640	0.0476
Shampoos	*Dove*	0.1085	0.0727	0.0577	0.0199
	Elseve	0.0619	0.0500	0.0241	0.0161
	Fructis	0.0501	0.0407	0.0188	0.0163
	Pantene	0.0841	0.0574	0.0449	0.0188
	Head & Shoulders	0.0891	0.0600	0.0468	0.0218

Table 5: Overview of spatial parameters

This table shows that for the more exclusive brands (i.e. “Scapa”, “Mercedes” and “Volvo”) the added value of the neighborhood variable reduces significantly on top of a traditional model in both models, while such a large drop of the spatial parameter estimates cannot be observed for the other public durable goods. For these brands, which are bought by a general public, it is more difficult to identify prospects only based on socio-demographic and lifestyle variables, resulting in a relatively poor traditional customer acquisition model (see Table 4). In such models, incorporating neighborhood effects can be very valuable to improve the identification of potential customers. Compared to public durable goods, the benefits of including spatial information is a lot smaller for privately consumed durable goods and, although still significant, very low for consumer packaged goods.

5. Discussion and Conclusion

Within customer relationship management, correctly identifying potential new customers can be a hard task because the information available is mostly limited to socio-demographic and lifestyle variables attracted from an external data vendor (Baecke & Van den Poel, 2011). In this context, augmenting these acquisition models with spatial information could improve the identification of prospects. Traditional CRM models often assume that customers act independently of each other, whereas in reality, the behavior of customers could be spatially correlated. In this case, it is preferable to use models that take advantage of this information instead of treating this as nuisance in the error term. This study applies two models (i.e. an autologistic model and a multilevel model) to investigate for 25 products and brands, divided over three categories, whether neighborhood effects could be identified and to what extent incorporating this spatial correlation can improve the predictive performance of customer acquisition models.

In a first step, the predictive performance of both spatial models is compared with a traditional CRM model. This comparison showed that both models are able to significantly improve the identification of customers across all of the 25 products and brands investigated in this study. When the predictive performance of both spatial models are compared with each other, both models perform equivalently when only spatial information is used as a predictor. However, this study finds that especially for durable goods, which are more exposed to neighborhood effects, a multilevel model is often better able to incorporate this spatial interdependence on top traditionally uses socio-demographic and lifestyle variables.

Further, this study also provides interesting insights for a marketing decision maker. Based on this comparison, involvement and visibility of a product turns out to be most determining whether neighborhood effects exist for a particular product or brand. Based on a model that only takes the spatial interdependence between customers into account, purchasing behavior is best predictable for public durable goods, followed by privately consumed durable goods. Predictions are worst for consumer packaged goods, which are not only privately consumed, but customers are generally also low involved with these products. Within each of the durable goods categories, it can be recognized that, next to involvement and visibility, also the exclusivity of the product has an influence on the amount of spatial interdependence. In other words, customers of more luxury product and brand (e.g. “Scapa”, “Mercedes”, “Volvo”, “Espresso machine”) are easier to be identified based on only spatial information. With these findings, this paper confirms based on a large behavioral data sample the surveyed result of Bearden & Etzel (1982) who found that publicly consumed luxuries are exposed to the most reference group influence.

However, remarkable is that although these luxuries experience the highest spatial interdependence, the model improvement is smaller than expected after the enhancement of a traditional customer acquisition model with spatial information. This is caused by the fact that these brands are often bought by a typical and more exclusive group of customers which are already easier to identify based on only socio-demographic and lifestyle variables. Further, the spatial variable can be a good proxy for these independent variables resulting in relatively high predictive performance of a model that is only based on spatial information. However, once this spatial variable is used in combination with socio-demographic and lifestyle variables, it loses a lot of his predictive power. In other words, although publicly consumed luxury durables are the most exposed to neighborhood effects, the augmentation of a customer acquisition model with spatial information is more valuable for products for which customers are difficult to be identified, such as more general, less exclusive brands.

Compared to publicly consumed durable goods, the added value of incorporating neighborhood effects is much more limited for privately consumed durables. For the identification of purchasers of specific CPG brands this added value is even smaller and, although still significant, economically less relevant. These findings are in line with the findings of Kuenzel & Musters (2007). Based on surveyed data, these authors found that also for low involvement products social influence can affect the purchase decision. However, this only exists between specific reference groups, such as close family or friends, but not between neighbors.

Based on 25 products and brands, this paper gives clear indications to marketing decision makers that spatial interdependence should not be neglected for certain types of goods. Instead of treating this as nuisance in the error term, taking advantage of this phenomenon can significantly improve a customer acquisition model. However, in order to generalize these findings, future research should examine even more product and brands. Besides this, it would be interesting to investigate whether incorporating spatial interdependence could also improve other CRM models, such as cross-sell, up-sell or churn models, which also includes transactional variables next to socio-demographic and lifestyle variables.

Acknowledgement

Both authors acknowledge the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

References

Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht: Kluwer.

Anselin, L. (2002). Under the hood Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*, 247-267.

Arndt, J. (1967). Role of Product-Related Conversations in the Diffusion of a New Product. *Journal of Marketing Research*, *4*(3), 291-295.

Augustin, N., Mugglestone, M., & Buckland, S. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, *33*(2), 339-347.

Baecke, P., & Van Den Poel, D. (2010). Improving Purchasing Behavior Predictions By Data Augmentation With Situational Variables. *International Journal of Information Technology & Decision Making*, *09*(06), 853-872.

Baecke, P., & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems*, *36*(3), 367-383.

Bearden, W. O., & Etzel, M. J. (1982). Reference Group Influence on Product and Brand Purchase Decisions. *Journal of Consumer Research*, *9*(2), 183-194.

Bell, D. R., & Song, S. (2007). Neighborhood effects and trial on the internet: Evidence from online grocery retailing. *Quantitative Marketing and Economics*, *5*(4), 361-400.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of Royal Statistical Society, Series B (Methodological)*, *36*(2), 192-236.

Besag, J. (1975). Statistical Analysis of Non-lattice Data. *Journal of Royal Statistical Society, Series D (The Statistician)*, *24*(3), 179-195.

Bradlow, E. T., Russell, G. J., & Bell, D. R. (2005). Spatial Models in Marketing. *Marketing Letters*, *16*(3-4), 267-278.

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9-25.

Bronnenberg, B. J. (2005). Spatial models in marketing research and practice. *Applied Stochastic Models in Business and Industry, 21*(4-5), 335-343.

Bronnenberg, B. J., & Mahajan, V. (2001). Unobserved Retailer Behavior in Multimarket Data: Joint Spatial Dependence in Market Shares and Promotion Variables. *Marketing Science*, *20*(3), 284-299.

Chen, W.-C., Hsu, C.-C., & Hsu, J.-N. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response model. *Expert Systems with Applications*, *38*(6), 7451-7461.

Courgeau, D., & Baccaini, B. (1998). Multilevel analysis in the social sciences. *Population: An English selection*, *10*(1), 39-71.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, *37*(3), 2132-2143.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837-845.

Du, R. E. X. Y., & Kamakura, W. A. (2011). Measuring Contagion in the Diffusion of Consumer Packaged Goods. *Journal of Marketing Research*, *48*(1), 28-47.

Grinblatt, M., Keloharju, M., & Ika, S. (2008). Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors. *The review of Economics and Statistics*, *90*(4), 735-753.

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45*(2), 171-186.

Hanley, J. A., & Mcneil, B. J. (1982). The meaning and use of area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

He, F., Zhou, J., & Zhu, H. (2003). Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(2), 205-222.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259-5264.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. New York: Taylor & Francis Group.

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299-310.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., et al. (2005). Choice Models and Customer Relationship Management. *Marketing Letters*, *16*(3-4), 279-291.

Kim, D., Lee, H., & Cho, S. (2008). Response modeling with support vector regression. *Expert Systems with Applications*, *34*(2), 1102-1108.

Kuenzel, J., & Musters, P. (2007). Social interaction and low involvement products. *Journal of Business Research*, *60*(8), 876-883.

Lee, V. E., & Bryk, A. S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education*, *62*(3), 172-192.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall.

Mcpherson, M., Smith-lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415-444.

Miller, H. J. (2004). Tobler ' s First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284-289.

Moon, S., & Russell, G. J. (2008). Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. *Management Science*, *54*(1), 71-82.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Detection Defection: Measuring of the Predictive Accuracy Understanding Models Churn Customer. *Journal of Marketing*, *43*(2), 204-211.

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602.

Pai, J.-C., & Tu, F.-M. (2011). The acceptance and use of customer relationship management (CRM) systems: An empirical study of distribution service industry in Taiwan. *Expert Systems with Applications*, *38*(1), 579-584.

Steenburgh, T. J., Ainslie, A., & Hans, P. (2003). Massively Categorical Variables: Revealing the Inforraation in Zip Codes. *Marketing Science*, *22*(1), 40-57.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, *39*(3), 2597-2605.

Wolfinger, R., & O'Connell, M. (1993). Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, *4*(3-4), 233-243.

Yang, S., & Allenby, G. M. (2003). Modeling Interdependent Preferences. *Journal of Marketing Research*, *40*(3), 282-294.

CHAPTER VI:

GENERAL SUMMARY, LIMITATIONS AND FUTURE RESEARCH

CHAPTER VI:
GENERAL SUMMARY, LIMITATIONS AND FUTURE RESEARCH

1. Introduction

During the last decades, there has been an important evolution within the field of marketing from mass marketing to database marketing and Customer Relationship Management (CRM). As a result of the technological revolution companies are able to personalize the interaction with their customers on a massive scale. As a result, the database of a company has become an increasingly important asset to support several marketing strategies such as customer acquisition, development and retention. This evolution has been reflected in the growing body of academic researchers, but also business practitioners, that have tried to improve these CRM models. In general, this can be done by focusing on one of two facets of CRM models. Firstly, a company can try to improve the methodology and data mining techniques used to obtain useful information from the available datasets. Secondly, a company can also try to enhance the database itself on which these data mining techniques are based. This last facet is the main focus of this dissertation. Each paper in this dissertation demonstrates how a traditional CRM model can be creatively augmented with new types of variables in order to improve the predictive performance of a CRM model. Table 1 gives an overview of the main predictive improvements as a result of enhancing a traditional CRM model with alternative variables. In each study, a logistic regression model is used as basic model for which predictive performance is calculated in terms of AUC as benchmark.

Chap- ter	Title	AUC Basic model	AUC Augmented model	AUC Improvement
II	Data Augmentation by Predicting Spending Pleasure using Commercially Available External Data [1]	80.45%	83.85%	+3.40%
III	Improving Purchasing Behavior Predictions by Data Augmentation with Situational Variables [2]	68.18%	70.35%	+2.17%
IV	Improving Customer Acquisition Models by Incorporating Spatial Autocorrelation at Different Levels of Granularity [3]	64.23%	67.83%	+3.60%
V	Including Spatial Interdependence in Customer Acquisition Models: a Cross-Category Comparison [4]	62.30% → 81.94%	62.44% → 82.42%	+0.14% → +2.18%

[1] Published in Journal of Intelligent Information Systems (2011)
[2] Published in International Journal of Information Knowledge and Decision Making (2010)
[3] In review for Journal of Intelligent Information Systems
[4] Published in Expert Systems with Applications (2012)

Table 1: Overview of predictive improvement.

Next, these models were augmented using commercially available external data (Chapter II), situational variables (Chapter III) or geographical data at multiple levels of granularity (Chapter IV). Chapter V examines the effect of incorporating geographical data across multiple product categories resulting in a wide range of predictive performance. In the last column of Table 1 the predictive improvement is presented, which is for each study not only statistically significant but also economically relevant. Steenburgh, Ainslie & Engebretson (2003) have illustrated that since such models are typically applied on a large number of prospects, even a small difference in predictive performance can lead to large differences in terms of profitability.

2. General summary

The first study, described in Chapter II, introduces a solution for the limitations of a company's database. These datasets are typically single sourced, containing only information about their own customers. Especially when a company would like to attract new customers this is a problem. Further, companies typically collect transactional information which only gives a view of the purchasing behavior of the customer, neglecting the attitudes that drive this behavior. In order to solve these limitations customers can attract data from an external data vendor. For such a data vendor, this study proposes a methodology to create variables that provide a solution for these limitations. These variables are created by linking surveyed data collected from a limited number of respondents with a large commercially available data set using predictive modeling techniques. This is based on a methodology similar to the one explained in the study of Lix, Berger, & Magliozzi (1995). However, this study uses a more advanced random forests technique that should avoid overfitting. This problem of overfitting can typically occur in this situation since a large number of variables is used to model the surveyed outcomes of a limited number of respondents.

However, in comparison to the study of Lix et al. (1995), the main contribution of this study is its focus on the construction of variables that should solve the limitation of a traditional database. Therefore, the surveyed variables are constructed in such a way that they do not only identify customers based on the purchase behavior in a particular product category, but also the attitude of the customer towards the product category is taken into account. Further, where Lix et al. (1995) mainly focused on the extrapolation methodology, this study also investigates whether these extrapolated variables are still powerful enough to improve an existing CRM model.

First of all, the results show that the effectiveness of linking surveyed data with a large database can differ significantly depending on the product category. The predictive performance, in terms of AUC, of the random forests technique used to execute these linkages ranges from 0.6375 to 0.8193. Hence, even the extrapolated variables at the lower end of this range can improve traditional CRM models significantly. This is demonstrated in an application in which a customer identification model of a magazine publisher is augmented with these variables. This application shows that on top of socio-demographic variables, such as age and gender, these extrapolated surveyed variables can assist in identifying potential customers. This has resulted in a lift of the predictive performance in terms of AUC from 80.45% for the basic model to 83.85% for the augmented model.

The second study, discussed in Chapter III, is the only study in this dissertation applied in the field of customer development. This study points out that the purchasing behavior of a customer does not only depend on the customer’s individual characteristics, but also situational variables can have an influence on the outcome of the purchase occasion. This has already been proven in a customer behavior study conducted by Belk (1975). Based on the theoretical framework of Belk (1975) this study investigates whether these finding can be exploited in predictive database marketing. For a home vending company, three dimensions of situational variables are included in a highly dynamical model used to predict the purchasing behavior of a customer on a daily basis. More specifically, this research investigates the augmentation of a traditional CRM model with situational information about physical surroundings, temporal perspective and social surroundings respectively represented by weather, time and salesperson variables. Since this last variable is a highly categorical variable that links each salesperson with a customer, a generalized linear mixed model is introduced, which allows for estimating a separate intercept for each salesperson.

A small but already significant improvement could be observed when the model takes the moment of the day the salesperson visits the customers into account. Furthermore, also weather information, such as "hours of sunshine", can be obtained in advance and improve the model significantly. The largest predictive improvement though, is measured by taking the salesperson effect into account. Finally, incorporating all three situational variables simultaneously on top of traditional transactional RFM variables resulted in a lift in terms of AUC from 68.18% for the basic model to 70.35% for the augmented model on the out-of-period test sample.

Chapter IV presents the third study in this dissertation, which focuses on the inclusion of geographic information in a CRM model. Traditional CRM models assume that the purchasing behavior of customers is totally independent. However, in reality, customer preferences do not only depend in their own characteristics, but are also influenced by the behavior of neighboring customers and their recommendations. Whereas in traditional CRM models neighborhood effects are treated as nuisance in the error term, an autoregressive approach can be applied to capture this interdependence and improve the predicted performance of the model (Yang & Allenby, 2003). Such a model includes a spatial lag effect by constructing a spatial weigh matrix that indicates the interdependent relationships between neighboring customers. Although the determination of interdependent relationships between customers is often very subjective, this could have an important influence on the model. Therefore, for a Japanese automobile brand, this study investigates the effect of the chosen granularity level, based on which the spatial weight matrix is constructed, on the predictive performance of the model. In addition, this study points out that the existence of spatial autocorrelation can have several origins. Beside social influence, also homophily and other exogenous shocks can cause correlating behavior between customers. This study suggests that this autocorrelation can be divided into several parts, each optimally measured on a different granularity level. Therefore, the autologistic model is extended to a model that simultaneously incorporates multiple levels of granularity. Finally, this study also investigated the effect of the sample size of a company's dataset on the optimal level of granularity.

Firstly, the results confirm the findings of Yang & Allenby (2003) that the incorporation of spatial interdependence can significantly improve the identification of prospects of a Japanese car brand. This improvement can be observed for each of the seven granularity levels examined in this study. Secondly, marketing decision makers should carefully consider the granularity level based on which the spatial weight matrix is constructed. This study proves that the proportion of predictive improvement heavily depends on the chosen granularity level. A spatial model based on a granularity level that is too coarse assumes dependence between too many customers that do not influence each other in reality. On the other hand, a model on a granularity level that is too fine could ignore correlation that exists in reality. Furthermore, also the stability of the spatial lag effect can be affected because the number of customers in each neighborhood becomes too small. Thirdly, this study suggests that attracting geographical information about customers on multiple granularity levels can be valuable. This information can then be used in a model that allows estimating each origin of spatial interdependence on its optimal granularity level. However, in a sensitivity analysis in which the effect of the sample size is examined, this study shows that such model is only useful when sufficient data is available. If the data is too limited, the spatial lag effects constructed on finer levels of granularity are based on too few observations. As a result, these spatial lag effects will not be stable enough to improve predictions. On the other hand, when a company is able to collect sufficient data, this research shows that such a model is significantly better in identifying prospects than the best performing “single level” autologistic regression model. Finally, compared to a traditional basic acquisition model, this model can improve the predictive performance in terms of AUC from 64.23%, for the basic model, to 67.83% for the augmented model that simultaneously incorporates multiple levels of granularity.

In Chapter V the last study of this dissertation is covered. This study also focuses on the incorporation of spatial interdependence in a customer acquisition model. This research paper contributes to previous research of Yang & Allenby (2003) and Steenburgh et al. (2003) in several ways. Firstly, each of these previous studies used a different model to take spatial interdependence into account. Whereas Yang & Allenby (2003) used an autoregressive approach, Steenburgh et al. (2003) demonstrated that also a hierarchical approach can incorporate a geographical variable that groups customers into mutually exclusive neighborhoods. In order to support marketing decision makers in choosing the most appropriate model, this study compares both approaches. Moreover, in order to generalize the finding that including spatial interdependence can significantly improve customer acquisition, this study uses both models in an application that tries to identify prospects for 25 products and brands.

When comparing the predictive performance of an autologistic model to a multilevel model, the results show that a multilevel model performs at least as well as an autologistic model. Especially when the impact of the neighborhood effects becomes more important (i.e., for durable goods), a multilevel model frequently outperforms an autologistic model. Further, when comparing the impact of this predictive improvement over several product categories, taking spatial interdependence into account is most valuable for publicly consumed durable goods, which are typically high involvement products. More remarkable is that for products exposing a relatively high amount of spatial interdependence in an empty spatial model, augmenting a traditional acquisition model with this information does not necessarily result in major predictive improvements. For these products, often luxury goods, the high spatial interdependence is mainly caused by homophily by which the spatial variable is a substitute for absent socio-demographic and lifestyle variables. As a result, these neighborhood variables lose a lot of predictive value on top of a traditional acquisition model that is typically based on such non-transactional variables. In contrast to previous studies, this research also investigates the effect of incorporating spatial information in CRM models for products with lower involvement. Similar to the consumer behavior study of Bearden & Etzel (1982), privately consumed durables are subject to less spatial interdependence resulting in a smaller predictive improvement. Besides durable goods, consumer packaged goods are examined because recently also for these goods reference group influences have been discovered (Du & Kamakura, 2011; Kuenzel & Musters, 2007). Although for these consumer packaged goods the predictive improvement is still significant in this study, the impact is quite small in order to be economically relevant.

3. Limitations and future Research

This dissertation provides valuable contributions to the existing marketing literature, and more specifically to the field of customer relationship management, by demonstrating the importance of data quality in CRM model. CRM improvements are not only obtained by improving the data mining techniques, but this dissertation illustrates that taking alternative data sources into account and incorporate these in a correct way can also be very profitable for a company. These days, company's datasets are often limited to traditional information such as socio-demographic variables, lifestyle variables and transactional data. Therefore, this dissertation provides several methodologies, divided over four studies, to create and/or incorporate alternative data, such as spending pleasure variables, situational variables and geographical data. Further, Table 1 illustrates that implementing these methodologies can result in significant predictive improvements, which will eventually result in better marketing strategies and higher company profits.

Despite these contributions, also several limitations can be identified in this dissertation, which entails opportunities for future research. The main limitation of this dissertation is that, except for the last study discussed in Chapter V, each study is based on a single dataset in a specific context. Therefore, in order to generalize the findings, future research should investigate whether similar conclusions could be made in other contexts. For example, in the first study, the benefit of extrapolating surveyed data using a commercially external database, as introduced by Lix et al. (1995), is examined in a customer acquisition context. Although this proved to be very valuable for optimizing customer identification, future research could investigate how such variables would act in a customer development or churn model, in which transactional variables are also included. The second study showed that by including the salesperson effect using a generalized linear mixed model the purchasing behavior could be significantly improved. These findings were found in a home vending context, but could also be tested in other contexts in which the link between salesperson and customer is easily identifiable, such as real-estate agents, investment advisers, insurance agents, etc.

Concerning the incorporation of spatial interdependence in CRM models though, previous research on this topic was also based on only single case studies (i.e. Steenburgh et al., 2003; Yang & Allenby, 2003). Chapter V tried to overcome this limitation by examining this topic on multiple product categories. However, also the two studies focused on the incorporation of neighborhood effects, discussed in Chapter IV and V, have some limitations. Firstly, both studies were applied in a customer acquisition context. Since these models are typically based on a limited amount of customer data, geographical information can quickly add value to the model. However, future research should investigate whether these variables would still be valuable on top of transactional information. Secondly, these studies use a variable that divides customers into mutually exclusive neighborhoods to measure spatial interdependence. Such variables, such as zip codes, are easily collectable and avoid the fact that observation on the edge of map would artificially have fewer neighbors in the model. Though, based on latitude and longitude variables, the real distance between customers could be incorporated. This allows for including even more variation in measuring neighborhood influences. Therefore, future research should investigate the effect of using a weight matrix based on the continuous distance between observations and compare this with the use of a contiguity matrix. Thirdly, this dissertation used spatial analysis only to incorporate information about the geographic proximity between customers, however, within marketing customers can also be mapped based on other characteristics. For example, Yang & Allenby, (2003) applied a spatial model, not only to include information from a geographically defined network, but also from a demographically defined network. Also Moon & Russell (2008) applied an autologistic model on a joint space map, which was created based on the past purchasing behavior of customers, to improve a product recommendation model. Future research should investigate whether these spatial analyses can be applied to other variables, such as lifestyle or social network variables.

Finally, this dissertation is mainly focused on data augmentation using offline data, namely surveyed data, situational variables and geographical variables. However, as shown in table 1 of Chapter I, previous research also indicates that data types such as web usage data, email interactions and network-based information are valuable data type to enrich a company’s database with. Especially since the internet becomes an increasingly important tool to improve the interactivity between company and customer, these online data types will gain importance within the field of data augmentation. Therefore, future research should elaborate also on these online types of data and investigate their value in comparison to the offline data types examined in this study. For example, social media is becoming very popular in our lives. Hence, the network data retrieved from these kinds of websites can be analyses in a similar way as done based on geographical data. Instead of assuming interdependence in purchasing behavior between customers living in the same neighborhood, also interdependence can be assumed between customers who are strongly linked on these social websites. In this way, it would be very valuable to investigate how these two types of interdependence differ from each other and to what extent these interdependences are overlapping.

References

Bearden, W. O., & Etzel, M. J. (1982). Reference Group Influence on Product and Brand Purchase Decisions. *Journal of Consumer Research*, *9*(2), 183-194.

Belk, R. W. (1975). Situational Variables and Customer Behavior. *Journal of Consumer Research*, *2*(3), 157-164.

Du, R. E. X. Y., & Kamakura, W. A. (2011). Measuring Contagion in the Diffusion of Consumer Packaged Goods. *Journal of Marketing Research*, *48*(1), 28-47.

Kuenzel, J., & Musters, P. (2007). Social interaction and low involvement products. *Journal of Business Research*, *60*(8), 876-883.

Lix, T. S., Berger, P. D., & Magliozzi, T. L. (1995). New customer acquisition: prospecting models and the use of commercially available external data. *Journal of Direct Marketing*, *9*(4), 8-18.

Moon, S., & Russell, G. J. (2008). Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. *Management Science*, *54*(1), 71-82.

Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively Categorical Variables: Revealing the Inforraation in Zip Codes. *Marketing Science, 22*(1), 40-57.

Yang, S., & Allenby, G. M. (2003). Modeling Interdependent Preferences. *Journal of Marketing Research*, *40*(3), 282-294.